ENCYCLOPEDIA OF

Physical Science and Technology

THIRD EDITION

Analytical Chemistry

AP

Table of Contents
(Subject Area: Analytical Chemistry)

Article	*Authors*	Pages in the Encyclopedia
Analytical Chemistry	*Ulrich J. Krull and Michael Thompson*	Pages 543-579
Atomic Spectrometry	*Vahid Majidi*	Pages 765-786
Auger Electron Spectroscopy	*C. L. Briant*	Pages 787-792
Capillary Zone Electrophoresis	*Tim Wehr*	Pages 355-368
Electrochemistry	*Donald T. Sawyer*	Pages 161-197
Electron Spin Resonance	*Larry Kevan*	Pages 331-345
Electrophoresis	*S. P. Spragg*	Pages 363-378
Elemental Analysis, Organic Compounds	*T. S. Ma*	Pages 393-405
Gas Chromatography	*Milos Novotny*	Pages 455-472
Infrared Spectroscopy	*Norman B. Colthup*	Pages 793-816
Liquid Chromatography	*Neil D. Danielson*	Pages 673-700
Magnetic Resonance in Medicine	*John F. Schenck*	Pages 959-981
Mass Spectrometry	*Kenneth L. Busch*	Pages 145-158
Mass Spectrometry in Forensic Science	*Jan Schuberth*	Pages 159-169

Microwave molecular spectroscopy	*Robert L. Cook*	Pages 799-852
Nuclear Magnetic Resonance (NMR)	*Bernard C. Gerstein*	Pages 701-720
Organic Chemistry, Compound Detection	*Raphael Ikan and Bernard Crammer*	Pages 459-496
Photoacoustic Spectroscopy	*Konka Veeranjaneyulu and Roger M. Leblanc*	Pages 1-13
Photoelectron Spectroscopy	*G. Hohlneicher and A. Gildenpfennig*	Pages 57-90
Raman Spectroscopy	*R. P. Van Duyne and C. L. Haynes*	Pages 845-866
Scanning Probe Microscopy	*C. Daniel Frisbie*	Pages 469-484
Sonoluminescence and Sonochemistry	*Kenneth S. Suslick*	Pages 363-376
Spectroscopy in Forensic Science	*Michael B. Eyring*	Pages 637-643
Thermal Analysis	*David Dollimore*	Pages 591-612
Tomography	*Z. H. Cho*	Pages 843-877
Ultrafast Spectroscopy and its Applications	*M. Hayashi, Y. M. Chang, T. K. Wang, S. H. Lin and*	Pages 217-226
X-Ray Analysis	*Ron Jenkins*	Pages 887-902
X-Ray Photoelectron Spectroscopy	*Charles C. Chusuei and D. Wayne Goodman*	Pages 921-938

Analytical Chemistry

Ulrich J. Krull
Michael Thompson
University of Toronto

I. Classical Methods
II. Instrumental Methods
III. Computers in Analytical Chemistry
IV. Future Perspectives

GLOSSARY

Accuracy Nearness of a measurement to its accepted value.

Analyte Species present in a sample of matter about which chemical information is sought.

Calibration Process of determining the precise quantitative relationship between a known concentration of a chemical species and a physical property of that species.

Interferences Elements or chemical compounds that have similar properties to the analyte that prevent its direct measurement.

Limit of detection Concentration of a chemical species that produces an analytical signal equal to twice the standard deviation of the background signal.

Matrix Matter present in a sample in which the species being determined is dispersed.

Noise Random fluctuations of analytical signal with time.

Precision Statistical measure of the distribution of a series of analytical determinations about the average value of the same analytical determinations.

Qualitative analysis Process of identifying one or more component(s) in a sample of matter.

Quantitative analysis Process of determining the relative amounts of one or more component(s) in a sample of matter.

Resolution The ratio given by the average magnitude of an analytical variable divided by the smallest difference of values of the analytical variable, where each value provides analytical signals that are considered distinct by statistical techniques.

Selective Method in analytical chemistry that yields a response for a group of chemical species.

Sensitivity Ratio of the change in the response of an instrument with a corresponding change in the concentration of a chemical species.

Specific Method in analytical chemistry that yields a response for a single chemical entity only.

Standard Chemical species with well-established physical properties that is employed to calibrate an analytical procedure.

ANALYTICAL CHEMISTRY is concerned with providing qualitative and quantitative information about the chemical and structural composition of a sample of matter. A huge variety of samples, from high concentrations of

elements in alloy steels to part-per-billion levels of drugs in biological tissue, are handled by the analyst. The field is founded on the conversion of a measured physical property of the species being examined to a usable signal. It is generally divided into two categories, classical and instrumental, on the basis of its historical development. The overall strategy is to prepare a sample correctly, choose a particular method of analysis, and report the results in a meaningful format, which may include a statistical evaluation.

I. CLASSICAL METHODS

A. Semimicro Qualitative Analysis

A complete system of qualitative analysis is usually restricted to the detection of a complex array of inorganic cations and anions. In principle, schemes are developed to include less common species such as rhenium and tellurate anions, but more often than not attention is concentrated on more common cations such as Na^+ and Cu^{2+} and oxy anions such as SO_4^{2-} and NO_3^-. The basis for identifying inorganic species lies in their chemical and physical behavior. For example, reagents are used to yield distinct chemical effects such as the production of colored solutions or precipitates, the generation of easily observed gases, and the dissolution of previously insoluble substances. In a typical analysis, the identification of several species is required, and in this circumstance, it is desirable to employ selective reactions for each component of the system.

This type of chemistry is incorporated into a strategy that involves the separation of the original mixture into several parts in order to avoid the buildup of a highly complex array with a specific sample due to the addition of a number of reagents. Each part is then subjected to an analysis of a small number of species. In summary, the analysis involves a set of sequenced separations and identifications.

The strategy for the separation of cations involves their division into a set of groups by treatment with a particular selective reagent. After solution of the unknown in water, a reagent is employed to cause all cations of the group to precipitate, with all other ions remaining in solution. A typical overall scheme (simplified) is depicted in Fig. 1. After *separation*, the precipitate is reexamined for specific

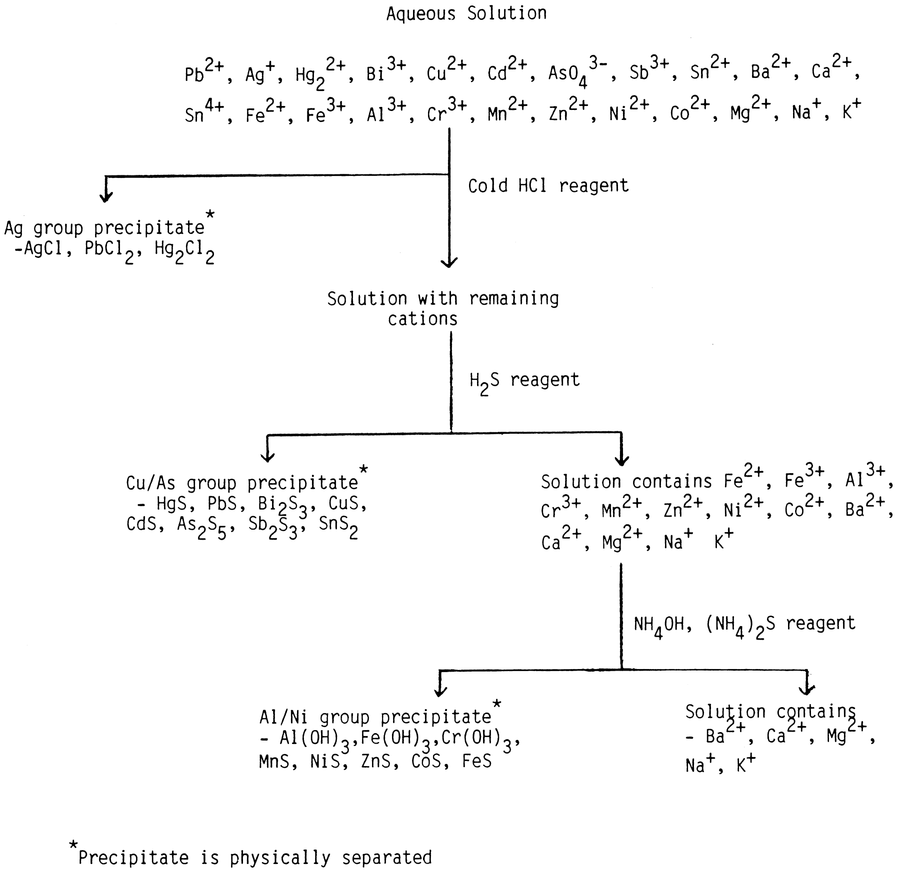


FIGURE 1 Simplified strategy for separation of cations into groups for qualitative analysis.

cations. As one example let us consider the silver group. Treatment of the white precipitate of Hg_2Cl_2, AgCl, and $PbCl_2$ with hot water results in solubilization of the lead compound to yield a colorless solution. The presence of Pb^{2+} is confirmed by the addition of K_2CrO_4, which gives a yellow precipitate of $PbCrO_4$. Addition of NH_4OH to Hg_2Cl_2 and AgCl results in either a gray mixture of mercury and $HgNH_2Cl$, confirming Hg_2^{2+}, and/or colorless solutions of $Ag(NH_3)_2^+Cl^-$. Acidification of the latter re-precipitates AgCl, identifying the presence of Ag^+.

The detection of anions in the original sample does not proceed in the same way in that the material is usually subjected to a series of preliminary tests. Moreover, the cation analysis described above can often be assessed to infer the presence of certain anions. A sample solution is separately treated with $AgNO_3$ and $BaCl_2$, and the various precipitates of silver and barium are used to verify the presence of anionic components. This procedure is then followed by the addition of concentrated cold H_2SO_4 to the unknown solid, which results in the liberation of characteristic gases (e.g., I^- gives the odor of H_2S and violet fumes of I_2), and by a series of confirmatory tests for each anion.

On a practical level the whole analysis is performed at the semimicro level. This means that sample sizes are in the range of 10 mg and solutions are of the order of 1–2 ml in volume. Special equipment is required to handle low volumes of reagent solutions and to avoid serious losses in separative and transfer procedures. The separation of precipitates is performed by centrifugation and decantation, and heating is achieved in specially designed test tube heating blocks.

B. Gravimetric Analysis

Here, quantitative analysis is based on the measurement of the weight of a substance of precisely known composition that is chemically related to the analyte. Most often the unknown is precipitated from solution by a reagent and, after separation and drying, is weighed. Less frequently the species being determined is volatilized, and the weight of the condensed gas or residual solid serves to complete the analysis. The precipitate must be insoluble in water (or other solvent) in order to minimize obvious losses, be easily filtered and washed, and be stable after drying or ignition procedures. With respect to the first requirement, an important quantitative parameter is the solubility product of the compound produced from the analyte. For example, for the precipitation of SO_4^{2-} with Ba^{2+}, the solubility product (K_{sp}) for $BaSO_4$ is defined by

$$K_{sp} = [Ba^{2+}]\left[SO_4^{2-}\right] = 1.3 \times 10^{-10}$$

for a saturated solution. Clearly, this value can be used to compute the loss of analyte (remaining in solution), and therefore the error, on reagent precipitation. The facility with which a precipitate is removed from solution is related to the particle size of the solid phase, which in turn is governed by the conditions that exist at the time of formation of the precipitate. Thought to be important are the solubility of the precipitate, temperature, reactant concentration, and mixing phenomena. These parameters control the supersaturation that exists in solution at a particular time. Although the usual aim is to achieve an easily manipulated crystalline precipitate, colloidal particles (10^{-6} to 10^{-4} mm in size) are sometimes obtained. In these cases, the individual particles must be coagulated by an experimental procedure such as heating, stirring, and the addition of auxiliary electrolyte. The mechanism of this process lies in the reduction of electrical repulsive forces on the particles. Purer, more dense precipitates can often be produced by precipitation from homogeneous solution, where the reagent is generated in solution. Finally, after filtration a gravimetric precipitate is heated until its weight becomes constant. A wide range of temperatures (110–1200°C) are used for this purpose, the most important requirement being weight constancy and unequivocal knowledge of the composition of the precipitate at a particular temperature.

Some examples of inorganic and organic precipitation agents for inorganic analytes are given in Table I. There are also methods available for the precipitation of organic compounds through the reaction of a particular functional group. For example, compounds with carbonyl functional groups can be precipitated with 2,4-dinitrophenylhydrazine according to the following reaction:

$$RCHO + H_2NNHC_6H_3(NO_2)_2 \rightarrow$$
$$R\text{—}CH{=}NNHC_6H_3(NO_2)_2 \downarrow + H_2O.$$

In summary, the gravimetric method does not require calibration, as is the case with many other analytical

TABLE I Typical Inorganic and Organic Precipitating Agents

Agent	Element precipitated (weighed form)
NH_3	Al (Al_2O_3), Fe (Fe_2O_3)
H_2S	Zn (ZnO), Ge (GeO_2)
H_2SO_4	Pb ($PbSO_4$), Ba ($BaSO_4$)
HCl	Ag (AgCl)
$(NH_4)_2CO_3$	Bi (Bi_2O_3)
$BaCl_2$	SO_4^{2-} ($BaSO_4$)
$MgCl_2$, NH_4Cl	PO_4^{3-} ($Mg_2P_2O_7$)
8-Hydroxyquinoline (HQ)	Al (AlQ_3)
Dimethylglyoxime (DMG)	Ni ($NiDMG_2$)

methods (see later). Accordingly, it is still frequently used as a "standardizing" technique for instrumental methods. Gravimetry, however, can be rather time consuming, especially if a large number of samples are involved.

C. Titrimetric (Volumetric) Analysis

In titrimetric analysis, which is often termed volumetric analysis, we obtain the volume of a standard reagent required to consume an analyte completely. On a practical basis a standard solution of reagent, the concentration of which is known accurately, is added by a buret until it is decided that the analyte is just used up. This condition is usually called the equivalence point. Since it is difficult to observe this point experimentally, it is usually approximated by the distinction of an end point, which is associated with detectable physical changes at equivalence. It is generally the case that standard solutions are calibrated against solutions of a primary standard that is a highly pure and stable reference substance.

In the present description we introduce four types of volumetric analysis. One involves the neutralization of acid or base, the second is concerned with precipitation, the third with complex formation, and the fourth with oxidation–reduction reactions. Before doing this we must consider several definitions. The equivalent weight of a substance is generally the weight that combines with a standard amount of reactant. The parameter is based entirely on the nature of a particular reaction and, therefore, can have no more precise meaning at this stage. Titration refers to the process of gradual addition of standard reagent, whereas the titer of a solution is the weight of a substance that is chemically equivalent to 1 ml of the solution. Finally, one often sees the term *normality*, which expresses the number of milliequivalents of solute contained in 1 ml of solution.

The end point in titrimetric methods is usually based on a way of dynamically distinguishing analyte or reagent concentration. The most widely used method is change in color due to reagent, analyte, or indicator, but the monitoring of electrical potential and current, and turbidity, are sometimes used.

1. Neutralization Titrations

In this technique the concentration of acid or base is determined through the abrupt change of pH that occurs at the titration end point. Here, the equivalent weight of the acid or base is the weight that either contributes or reacts with 1 mol of hydrogen ion in that specific reaction. The standard reagents used in the titration are always strong acids or bases, although the analyte may be a weak acid or base. A-standard acid solution can be prepared by titration of the acid, say HCl, against a solution containing an accurately known weight of Na_2CO_3. In a similar fashion, bases are standardized against potassium hydrogen phthalate ($KHC_8H_4O_4$). The end point in a strong acid–strong base neutralization titration is usually found from the *in situ* behavior of an added indicator, which is generally a weak organic acid or base that undergoes chemical changes exhibiting different colors. For example, we can write for the acid-type indicator HIn

$$H_2O + HIn = H_3O^+ + In^-,$$

(color 1 in acid solution) (color 2 in basic solution)

Thus, it is very important to know the range of hydrogen ion concentration (i.e., pH) in which a change from color 1 to color 2 can be observed. Generally, this occurs within approximately ±1 pH unit of the pK_a of the indicator.

The titration of a strong acid solution by additions of a standard strong base solution from a buret would result theoretically in the lower graph shown in Fig. 2 (the titration curve). Accordingly, we would require an added indicator to respond to the abrupt approximately 3–10 pH change. In this case, phenolphthalein would be appropriate, since its change from colorless to purple is easily detected by the eye. The titration curve for a weak acid (with strong base) depends very much on the dissociation constant of the acid. Basically, additions of base set up a buffer zone as in the set of curves shown in Fig. 2 (upper curve). The choice of an indicator for this type of titration is significantly more limited than for a strong acid, since there is now no abrupt change in pH. Not surprisingly, the appropriate curves for polyprotic weak acids are even more complex.

Neutralization methods are employed wherever inorganic or organic substances possess acidic or basic groups. An important application is the conversion of elements

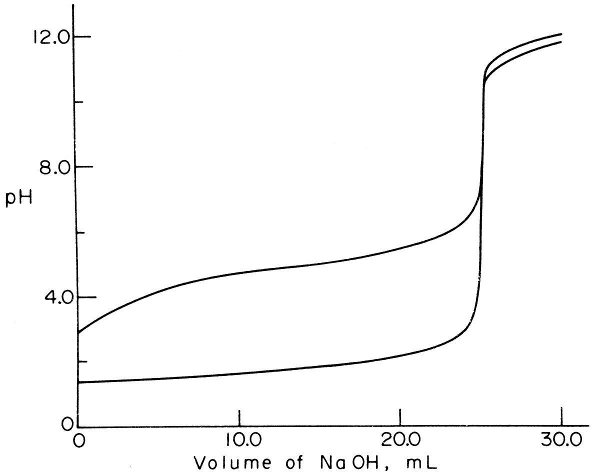


FIGURE 2 Neutralization titration curves for 50 ml of 0.05 *M* strong acid (lower curve) and weak acid (upper curve) with 0.1 *M* NaOH solution.

in organic or biological systems to acid–base character. For example, organic nitrogen is converted to NH_4^+ in the Kjedahl method to be determined in a neutralization procedure as liberated NH_3.

2. Precipitation Titrations

In this method, the reagent solution, added from a buret, is allowed to cause precipitation with the analyte, and the end point is associated with an abrupt change in the concentration of the *reagent* species. The equivalent weight of a participant in a precipitation titration is the weight that reacts with or provides 1 g formula weight of the reacting cation if it is univalent (0.5 for divalent, etc.). The technique can be best understood by the example provided by the long-standing silver ion procedure for the determination of halide anion, particularly Cl^- (sometimes termed argentometric titration). Typical titration curves for various halide anions with silver cation are shown in Fig. 3. Notice that the concentration of the free reagent cation present at a particular point in the titration is usually expressed as $-\log[Ag^+]$, that is, pAg, in a fashion analogous to pH for hydronium ion. Before the equivalence point (with abrupt change of pAg), the concentration of Ag^+ is determined by the solubility product of the silver halide concerned. After this point we simply have the concentration of excess added reagent, which is not now reacting with halide anion. To detect the end point an indicator is required that will respond to the particular range of abrupt pAg change shown in Fig. 3. A common method in the case of Cl^- and Br^- is that of Mohr, who used chromate ion (CrO_4^{2-}), the end point being specified by the appearence of Ag_2CrO_4, which is brick red in color. The essence of the principle here is that the solubility of silver chromate is much greater than that of silver halide. Accordingly, the conditions of the titration can be adjusted such that the formation of silver chromate occurs only in the vicinity of the end point. In the Volhard method a standard solution of thiocyanate ion (SCN^-) is titrated against Ag^+ using Fe^{3+} as the indicator. At the initial excess of SCN^-, a red color due to red $Fe(SCN)^{2+}$ is produced. An important application of this method is the "backtitration" procedure for the Cl^- ion. Here, an excess of standard silver nitrate solution is added to the unknown Cl^- solution and then titrated with SCN^- with the Volhard indicator in place. Finally, adsorption indicators (Fajans method) are employed that are based on the adsorption of a colored organic compound to the precipitate, *only* in the vicinity of the end point.

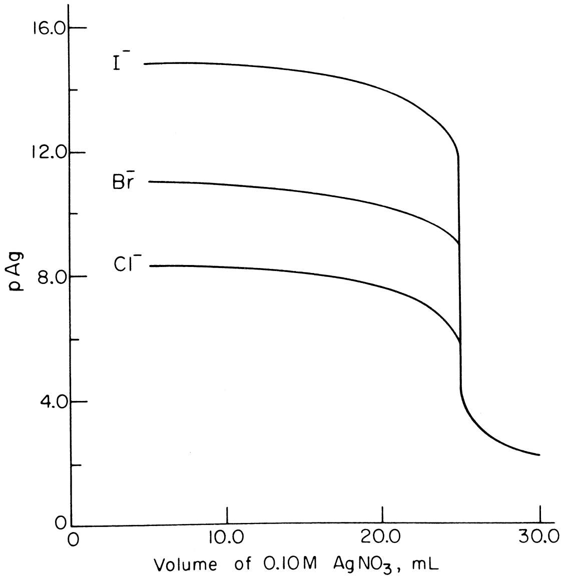
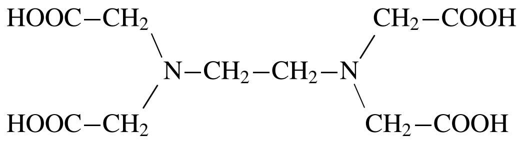


FIGURE 3 Precipitation titration curves for 25 ml of 0.1 *M* solutions of various anions with 0.1 *M* $AgNO_3$ solution.

3. Complexometric Titrations

Titrimetric methods based on the reaction of metal ions with a coordinating species (ligand) are more than 100 years old. If the ligand is attached to the metal by more than one functional group, it is said to be polydentate and the compound produced is called a chelate. The complexometric reagent that is easily the most widely used for the titration of unknown concentrations of metal ion is the hexadentate ligand ethylenediaminetetraacetic acid (EDTA), first recognized by Schwarzenbach in 1945. The structure of the molecule is



and is usually abbreviated to the shorthand notation H_4Y, with H representing the carboxylic hydrogen atoms. This reagent is extremely important because it forms simple 1 : 1 complexes with a metal ion, that is, one metal atom to one EDTA molecule of very high stability. Note that the latter is associated with the relatively large number (6) of points of attachment ($4 \times COO^- + 2 \times N$) of the ligand to the metal.

In the now familiar pattern discussed above, the titration involves the buret addition of EDTA solution to the metal ion solution, which generates a titration curve with an abrupt change in $-\log[M^{n+}]$ (pM). This is governed by the equilibrium constant for the formation of the metal–EDTA complex:

$$K_{MY} = \frac{[MY^{(n-4)+}]}{[M^{n+}][Y^{4-}]},$$

Note the analogy of this result with that outlined for the effect of solubility product in precipitation titrations discussed above. A great many compounds have been proposed as indicators for metal ions in EDTA titrations. These species are generally organic compounds that form colored chelates with metal ions in a range of pM that is characteristic of the cation and dye. One example is Eriochrome black T, which is blue at pH 7 and red when complexed with a variety of metal ions.

EDTA titrations are still widely used because of their great versatility with respect to the analysis of a large number of different metal cations. Furthermore, the technique can be made more selective by adjusting the pH or by the use of compounds that effectively remove interfering cations from the titration (masking agent). The method is inexpensive and reasonably accurate.

4. Oxidation–Reduction Titrations

Here, we are dealing with the reaction of an oxidizing (reducing) reagent as titrant with a reducing (oxidizing) species as unknown. The equivalent weight of a participant in this type of system is the weight that directly or indirectly produces or consumes 1 mol of electrons. Thus, the equivalent weight for the permanganate ion (MnO_4^-) in its oxidation of oxalate anion in the reaction

$$5C_2O_4^{2-} + 2MnO_4^- + 16H^+ \rightarrow 10CO_2 + 2Mn^{2+} + 8H_2O$$

is the molecular weight of MnO_4^- divided by 5.

A great variety of both oxidizing and reducing agents have been employed for this type of titration, and some common compounds are given in Table II. The equivalence point of oxidation–reduction titrations can be computed from a knowledge of solution concentrations and electrical potentials. For example, in the titration of ferrous ion against $KMnO_4$ according to the following reaction

$$5Fe^{2+} + MnO_4^- + 8H^+ \rightarrow 5Fe^{3+} + Mn^{2+} + 4H_2O,$$

we can write Nernst electrode potentials for each oxidation–reduction system:

$$E = E^0_{Fe^{3+}} - 0.0591 \log \frac{[Fe^{2+}]}{[Fe^{3+}]}$$

$$E = E^0_{MnO_4^-} - \frac{0.0591}{5} \log \frac{[Mn^{2+}]}{[MnO_4^-][H^+]^8},$$

At the equivalence point we know that these two potentials, now called E_{eq}, are equal; therefore, on adding we derive

$$E_{eq} = \frac{E^0_{Fe^{3+}} + 5E^0_{MnO_4^-}}{6} - \frac{0.0591}{6} \log \frac{5[MnO_4^-][Mn^{2+}]}{5[Mn^{2+}][MnO_4^-][H^+]^8}.$$

Thus,

$$E_{eq} = \frac{E^0_{Fe^{3+}} + 5E^0_{MnO_4^-}}{6} - \frac{0.0591}{6} \log \frac{1}{[H^+]^8}.$$

TABLE II Common Oxidizing and Reducing Agents Employed as Standard Solutions in Oxidation–Reduction Titrations

Reagent	Oxidizing species	Reduction Product
Oxidizing		
Potassium permanganate, $KMnO_4$	MnO_4^-	Mn^{2+}
Potassium bromate, $KBrO_3$	BrO_3^-	Br^-
Cerium ammonium nitrate, $Ce(NO_3)_4 2NH_4NO_3$	Ce^{4+}	Ce^{3+}
Potassium dichromate, $K_2Cr_2O_7$	$Cr_2O_7^{2-}$	Cr^{3+}
Potassium, iodate, KIO_3	IO_3^-	I^-
	Reducing species	**Oxidation product**
Reducing		
Sodium thiosulfate, $Na_2S_2O_3$	$S_2O_3^{2-}$	$S_4O_6^{2-}$
Ferrous ammonium sulfate, $Fe(NH_4)_2(SO_4)_2{\cdot}6H_2O$	Fe^{2+}	Fe^{3+}

In much the same manner as for the other types of titration described above, it is the electrode potential during the procedure that exhibits an abrupt change on addition of reagent. Accordingly, we require indicators that change color during this change (i.e., that show oxidation–reduction behavior themselves). Two examples are 1, 10-phenanthroline–iron (II) complex, which changes from pale blue to red at an electrical potential of +1.11 V (in 1M H_2SO_4), and diphenylaminesulfonic acid, which changes from colorless to violet at 0.85 V.

II. INSTRUMENTAL METHODS

A. Absorption of Electromagnetic Radiation

Matter interacts with incident electromagnetic radiation by the three distinct processes of transmission, scattering, or absorption. The nature of any interaction is a function of the properties of the radiation, such as energy, phase, polarization, and the chemical properties of the matter under

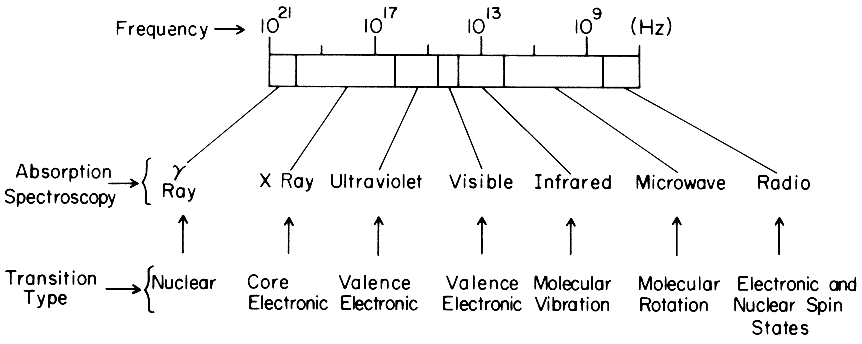


FIGURE 4 Analytical potential for absorption spectroscopy across the electromagnetic spectrum.

investigation. Chemical evaluation is derived from observation of the extent of interaction by measurement of the energy and intensity of transmitted, scattered, absorbed, or latently released radiation.

Absorption is said to occur when radiation passes through matter and interacts with the ions, atoms, or molecules constituting the sample in such a way that they gain energy and move from a quantized low-energy state to a higher energy state. Analytically, this phenomenon is observed as a reduction of the intensity of radiation after passage through a sample of matter. Two distinct parameters can be investigated. The reduction of intensity is a function of the probability of interaction of the radiation with appropriate analyte species and indicates concentration. Also, the energy of the absorbed radiation indicates the quantized energy levels in which energy is deposited (Fig. 4) and therefore assists in species identification. The Beer–Lambert law generally applies to all absorption processes at low concentrations when monochromatic radiation is used. It is conventionally written

$$\log\frac{P_0}{P} = \varepsilon bc = A,$$

where P_0 is the initial power of an incident beam of radiation, P is the final power (decreased due to absorption), ε represents a value characteristic of the extent of absorption expected for a certain species at a defined wavelength in a defined matrix, b is the sample path length, c is the concentration of the analyte, and A represents a value for absorption.

1. Atomic Absorption

The absorption of radiation by a sample of atomic particles, created by vaporizing the sample, represents a relatively simple spectral situation that has great practical value for elemental identification and concentration determination. The absorption spectrum observed with polychromatic light contains only a few areas of reduced intensity occurring at very well-defined frequencies due to the small number of energy states available to the atoms. The natural width of such absorption lines can be less than 10^{-4} nm, but broadening often occurs due to collisional and Doppler effects in the sample matrix. The quantized energy transitions responsible for atomic absorption are of electronic origin, indicating that electromagnetic spectrum energies from X-rays to ultraviolet–visible rediation are sufficient to observe this process.

The technique known as atomic absorption spectroscopy is of particular analytical importance for the determination of metals due to its sensitivity and potential for selectivity by virtue of the narrow atomic absorption lines. A schematic representation of the spectrophotometer is shown in Fig. 5. To take advantage of characteristic selectivity, a special radiation source must be provided to produce extremely monochromatic radiation with a

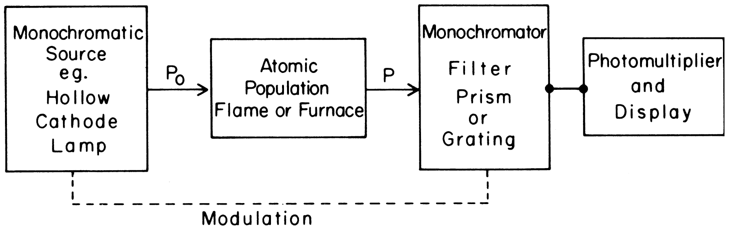


FIGURE 5 Representation of the design of a conventional atomic absorption spectrophotometer.

linewidth less than the absorption bandwidth of the atomic sample. In most cases the same element being analyzed is also used in the source lamp to provide appropriate emission line spectra by stimulation with a plasma discharge. Common sources are hollow cathode lamps, electrodeless discharge lamps, and gaseous discharge lamps. The source is often modulated and coupled to a lock-in amplifier in order to determine the relative absorbance of the atomic sample as compared with sample emission at the same wavelength. Other methods of background correction involve monitoring of a nonabsorbed radiation by the use of distinct spectral lines, continuous sources, or the Zeeman effect (separation of degenerate energy states of an atom by application of a powerful magnetic field; provides closely spaced energy levels which are sensitive to polarized light). The sample itself is usually produced by volatilization of solutions in a high-temperature flame or solids and in a graphite furnance. Flame methods generally provide relative errors of approximately 1–2%, whereas the equivalent furnace methods provide only 5–10%, but sensitivities up to 1000 times greater than those observed for flames. Chemical interferences commonly originate from the atmosphere supporting the atomic population and from other sample matrix components. The formation of oxides in flames causes the reduction of the atomic population and can be controlled by the employment of appropriate fuel–oxidant ratios and by sampling appropriate portions of the flame. This problem is avoided in the furnace by the use of a continuous flowing inert gas atmosphere. Poor volatilization and atomization often occur when an analyte binds with anions in the sample matrix. This is overcome by the use of high concentrations of cationic releasing agents, which preferentially couple with the interferent, or by the use of protective agents such as chelating ligands, which form stable but volatile species with the analyte of interest. Spectral interferences can also occur, though direct overlap of interferent and analyte atomic absorption bands is rare. These interferences include molecular band absorption such as that experienced from the formation of combustion products, which can often be eliminated by the use of higher temperatures, and ionization of the atomic population due to excess temperatures. Since many of the chemical processes occurring in certain localized areas of the atomic atmosphere are approximately in equilibrium, such ionization can be suppressed by the addition of a radiation buffer, which selectively ionizes in contrast to the analyte of interest and therefore increases the probability of ion–electron recombination to form the required atomic population. Table III presents some of the analytes commonly investigated by the techniques of atomic absorption spectroscopy.

TABLE III Atomic Absorption Detection Limits for Common Analytes

Element	Analytical wavelength (nm)	Limit of detection (ppm)	
		Flame absorption[a]	Nonflame absorption
Aluminum	396.2	0.03	0.00001
Calcium	422.7	0.001	0.00005
Cadmium	326.1	0.0005	0.000003
Iron	372.0	0.003	0.00002
Lithium	670.8	0.0005	0.0003
Magnesium	285.2	0.0001	0.000004
Potassium	766.5	0.005	0.0009

[a] Data for acetylene–air flame.

2. Molecular Absorption

The absorption spectrum commonly produced by molecules is significantly more complex than that produced by atoms due to the large number of energy states available for energy deposition. The total energy of a molecule that has no translational velocity is given by

$$E_{\text{total}} = E_{\text{electronic}} + E_{\text{vibrational}} + E_{\text{rotational}},$$

where the electronic levels due to electrons in bonding and nonbonding orbitals are related to numerous interatomic vibrations and molecular rotations about a center of gravity. The large energy difference of the three energy types implies that each can be individually studied with radiation from distinctly different portions of the electromagnetic spectrum.

Optical polarization. Since the early 1800s it has been recognized that certain molecules have the capacity to rotate plane-polarized light, which has since led to the development of a number of techniques suitable for qualitative structural determination and quantitative concentration analysis. All these techniques are based on the fact that electromagnetic radiation has wavelike properties and can be represented as a combination of electric vectors. The electric vector can interact with the electrons of matter in an absorption–reemission process taking place over a time period of 10^{-14} to 10^{-15} sec. Even though this process does not change the energy of the radiation, a slowing does occur. If a monochromatic beam of radiation passes through an anisotropic solid sample, the electric vectors that encounter greater particle density will be slowed more than the vectors passing through less dense areas. Plane-polarized light is represented as the resultant of two interfering electric vectors such that the resultant always lies in a single plane. In the case of plane-polarized monochromatic light with two coherent perpendicular electric vector

components, the components can pass through the sample and be unequally retarded so that the recombination after exiting the sample results in constructive interference with a phase difference from that originally observed. This results in a rotation of the plane of polarization and propagation of circularly or elliptically polarized light dependent, respectively, on phase separations of multiples of 0, $\frac{1}{4}$, $\frac{1}{2}$, or combinations of these numbers.

Polarimetry is a technique suitable for structural determination and quantitative analysis and simply studies the direction and degree of the rotation of plane-polarized monochromatic radiation of transparent substances. Instruments generally make use of vapor lamps, which isolate the sodium D line or mercury line at 546 nm. The light is then processed by a Nicol polarizer prism assembly, passed through a half-shadow Lippich prism (to assist detection of the optical minimum intensity), passed through a thermostated transparent sample holder, passed through a second Nicol prism known as the analyzer, and directed to an eyepiece or photoelectric detector. Rotation of the analyzer allows matching of the split-field bright image. The instrument is calibrated by analyzing the rotation of a standard solution of known concentration and rotatory power and provides precision to 0.001° when coupled to photoelectric detectors.

Other techniques that find widespread application are optical rotatory dispersion and circular dichroism. These are closely related methods that rely on the wavelength dependence of refractive index and molar absorptivity, respectively, toward the direction (d or l) of circularly polarized radiation. Optical rotatory dispersion experiments provide a measurement of molecular rotations as a function of wavelength. Circular dichroism results are presented in a plot of wavelength vs a value $[\theta]$, and $[\theta]$ is a corrected value of θ, the molecular ellipticity measured as the singular difference between the emergent elliptical beam and the original incident plane of polarization. These techniques have been important tools in the elucidation of structural information about complex organic and biological systems and metal–ligand complexes.

The use of elliptically polarized light has also found application in the determination of refractive index and thickness of ultrathin films. Circularly or elliptically polarized light is reflected from a surface on which the sample is coated. The reflection through the sample alters the polarization, and the degree of alteration provides thickness data to 0.1-nm precision. These instruments have wide applicability in the semiconductor industry and as tools for the measurement of organic films of monolayer dimensions.

Ultraviolet–visible. This form of electromagnetic radiation is of sufficient energy to cause electronic absorption transitions from the ground state to an excited state within molecules. In contrast to sharp atomic electronic absorptions, molecular spectra show broad bands covering tens of nanometers due to the availability of a multitude of distinct energy states caused by the overlap of the vibrational and rotational energy sublevels with the primary electronic level. Only high-resolution spectrometers can clearly resolve these energy levels.

The design of a typical double-beam instrument that allows for relative background correction is schematically shown in Fig. 6 and consists of a radiant source, such as a hydrogen–deuterium lamp for ultraviolet and a tungsten lamp for visible radiations; a filter, prism, or grating monochromator for wavelength selection; sample and reference cells; and a photocathode, usually associated with a photomultiplier for detection.

The most important types of electronic transition involve the following:

1. n, σ, and π electrons, covalent compounds
2. Charge-transfer electrons
3. Electrons, ligand field strength
4. f electrons, lanthanide and actinide ions

Of particular importance are those transitions involving bonding electrons, since particular absorption energies are characteristic of the bonding of certain functional groups, known as chromophores, providing chemical identification and concentration information. A summary of approximate electronic energy transitions for bonding and nonbonding electrons is given in Fig. 7. This can be correlated

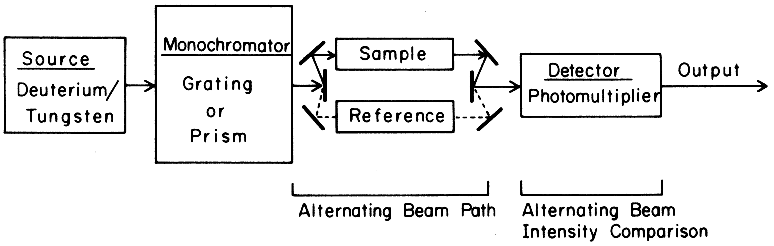


FIGURE 6 Conventional double-beam ultraviolet–visible absorption spectrophotometer.

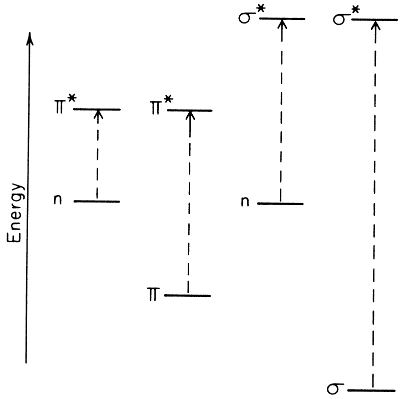


FIGURE 7 Electronic transitions for *n*, σ, and π electrons caused by ultraviolet–visible energy absorption.

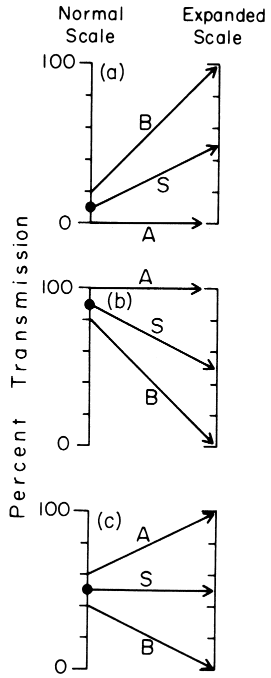


FIGURE 8 Differential absorption methods for quantitative concentration analysis. (a) High-absorbance method: A, optical shutter; B, low-transmission reference solution. (b) Low-absorbance method: A, pure solvent; high-transmission reference solution. (c) Ultimate precision method: A, B, reference solutions; S, sample.

with the wavelength for maximum absorption of some typical chromophores listed in Table IV. Molecular structural information can occasionally be garnered from wavelength shifts for particular chromophores when they are affected by conjugation or by structural distortion. Although all covalently bonded electrons can be excited to higher energy levels, the energy associated with single-bond transitions lies in the vacuum ultraviolet range, which because of technical difficulty is usually not exploited. This implies that analyses of unsaturated species and aromatic compounds are by far the most prevalent.

Due to the broadband nature of the electronic absorptions, this technique is poor for qualitative information acquisition but is excellent for accurate quantitative analysis. Three differential absorption methods are employed to maximize the accuracy of the analysis, with the choice being dependent on the concentration of the absorbing species and the availability of appropriate calibration standards. The three methods are outlined in Fig. 8.

TABLE IV Absorption Characteristics of Common Ultraviolet–Visible Chromophores

Unconjugated chromophore ($n \to \pi^*$ transitions)	Solvent	Approximate λ_{max} (nm)[a]	Approximate ε_{max}
Aldehyde	*n*-Hexane	293	12
Alkyne	*n*-Heptane	225	160
Amide[b]	Vapor phase	215	600
Amido	Water	214	60
Carboxyl	Ethanol	204	41
Ketone	*n*-Hexane	280	16
Nitrate	Dioxane	270	12
Nitro	Isooctane	280	22

[a] Considering only conventional instruments scanning λ greater than 200 nm.

[b] Amide absorption at this λ_{max} due to $n \to \sigma^*$ transition.

Infrared. It can be seen from Fig. 4 that infrared radiation is of an energy suitable for the stimulation of vibrational absorption that occurs within a single electronic level. For gas-phase samples, vibrational bands can be divided into a series of closely spaced absorptions associated with molecular rotational energy levels; however, these are not usually observed. Even so, the infrared absorption spectrum for most solid and liquid samples is complex and contains many sharp maxima and minima associated with well-defined vibrational energy levels.

Vibration consists of the periodic fluctuation of atoms with respect to relative position, but not all such position changes are associated with radiation absorption. Only when a net molecular dipole change is associated with vibrational or rotational motion can the alternating electric field of the radiation interact with the matter. Absorption then leads to a resonant vibrational or rotational amplitude increase.

All vibrations can be classified into the two major categories of stretching, where bond lengths alternate, or bending, where angular changes occur between two bonds. Theoretically and generally, the vibrational modes, amplitudes, and energies can be approximated by classical and quantum mechanical treatments of a harmonic oscillator at low potential energies. Given that the description of any atom in space requires the use of three coordinates

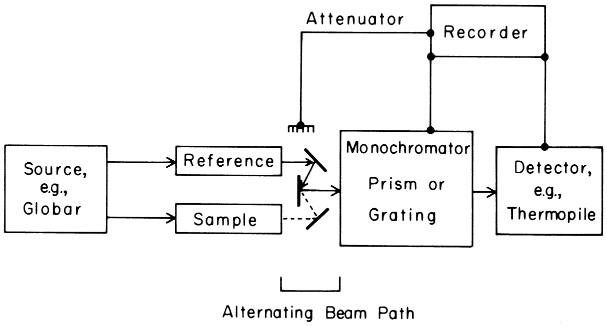


FIGURE 9 Conventional double-beam infrared absorption spectrophotometer.

to locate position and a molecule contains N atoms, then $3N$ coordinates, or "degrees of freedom," are required to describe the molecule. Since molecular motion consists of translation, vibration, and rotation and three coordinates are required to describe translation and also rotation, $3N - 6$ degrees of freedom remain to describe the number of possible "normal" modes of vibration. The number of normal modes of vibration does not necessarily correspond to the total number of observed vibrational absorptions, since some extra vibrational signals can be gained from overtone and combination frequencies, while some can be lost by being dipole inactive, being outside the instrumental analysis range, by overlap, or by having too low an intensity.

The design of a typical double-beam infrared spectrophotometer is schematically illustrated in Fig. 9. Typically, the broadband source consists of a metal wire or ceramic tube heated to incandescence by passage of an electric current. The radiation is first divided into two beams, which are directed through the sample and reference cells. A chopper mechanism placed behind the cells alternately selects transmission of either the sample or reference beam to the grating monochromator and thermal detector. This sets up an alternating current (ac) output from the detector, which is passed to a synchronous rectifier. A comparison of beam power occurs via the rectifier, which is produces a continuous unfluctuating direct current (dc) if the beams are identical. If the beam powers differ, an ac current is output from the rectifier, and after further amplification, this output signal is used to drive a synchronous motor. The motor concurrently drives a recording pen and an attenuator, until the attenuator causes the two beams to be of equal power and a signal null is achieved.

Though numerous instruments and measurement methods have been devised for quantitative work, the large number and narrow characteristics of infrared absorption bands make them more suitable for qualitative identification. This is especially true of the region between 2000 and 750 cm^{-1}, which is often called the "fingerprint" region. Such analyses can be applied to solid-, liquid-, and gas-phase samples, and a summary of some common absorption bands useful for chemical identification purposes is provided in Table V.

Electron spin resonance. In a strong magnetic field, the degenerate energy levels designated by the electron spin quantum number $\pm\frac{1}{2}$ actually differ in energy. The difference in energy between these levels is described as ΔE,

$$\Delta E = \mu\beta_{\text{N}} H_0 / I,$$

where μ is the magnetic moment, β_{N} the Bohr magneton, H_0 the external magnetic field strength, and I the quantum spin number. The difference in energies between the two distinct states can be equal to that of a photon in the microwave region of the electromagnetic spectrum and, for a conventional magnetic field strength of 3400 G, represents a frequency of 9500 MHz. It is possible for energy absorption to occur, promoting an electron from the low-energy state to the high-energy state by capture of photons in the microwave region. The majority of molecules do not exhibit an absorption spectrum since all electrons are paired and equal numbers exist in the two spin states. Paramagnetic molecules such as free radicals are strongly influenced by magnetic fields. The associated splitting of energy levels is very evident from the presence of absorption bands, which may be complicated by hyperfine splitting caused by electron spin–nuclear spin coupling.

A typical electron spin resonance instrument consists of a microwave source, known as a klystron tube, which by

TABLE V Infrared Absorptions of Common Chromophores

Chromophore	Wavelength range (μm)	Vibration mode and intensity
Alcohols	3.1–2.7	O—H stretch (strong)
Amines	3.3–2.8	N—H stretch (medium)
C—H	3.8–3.0	C—H stretch (strong)
Cyanides	4.7–4.4	C≡N stretch (medium)
Alkynes	5.1–4.6	C≡C stretch (weak)
Carbonyls	6.5–5.5	C═O stretch (strong)
Alkenes	6.4–5.9	C═C stretch (variable)
N═O	6.8–6.1	N═O stretch (strong)
Amides	8.4–6.0	N—H deformation (medium)
Esters	8.6–7.6	RCOOR (variable)
Alcohols	10.0–8.5	C—OH stretch (strong)
C—Cl	15.4–12.5	C—Cl stretch (strong)
C—Br	17.9–13.3	C—Br stretch (strong)

means of a metallic waveguide transmits the electromagnetic radiation to a quartz tube sample holder positioned between the poles of a permanent magnet. The experiment consists of varying the magnetic field strength by means of secondary coils to induce resonance. The resulting absorption signal is usually processed and outputed as a first-derivative curve to increase sensitivity and resolution. Though the technique does not have wide applicability, it continues to play an important role in the probing of biological systems for order, viscosity, reactivity with spin-label reagents (since absorption is influenced by chemical environment), and investigation of chemical reactions, which include free-radical intermediates.

3. Nuclear Absorption

Nuclear magnetic resonance. Certain atomic nuclei have properties of spin and magnetic moment that cause their energies to split into quantized levels when subjected to a powerful magnetic field. Transitions from a lower energy level to a higher level can be induced by absorption of electromagnetic radiation in the radiofrequency range of 0.1–100 MHz. Such absorption can occur only when the energy provided is equivalent to the energy difference between two magnetically induced states and satisfies the resonance condition. The experiment that measures the frequencies that satisfy the absorption process is known as nuclear magnetic resonance spectroscopy. This technique is related to electron spin resonance, which has been described previously.

For a nucleus of spin quantum number I, a total of $2I + 1$ discrete energy states exist in a magnetic field. If u represents the magnetic dipole caused by the spin of the charged nucleus, the energy difference ΔE between states can be written generally as

$$\Delta E = \mu\beta H_0/I,$$

where H_0 is the magnetic field strength and β is a constant called the nuclear magneton. A term that is characteristic of a nucleus is known as the magnetogyric ratio γ,

$$\gamma = \mu\beta/I(h/2\pi),$$

where h is Planck's constant. This value can be used directly to relate the frequency of absorbed radiation v to the magnetic field strength by the relation

$$v = \gamma H_0/2\pi,$$

Some nuclei commonly observed by nuclear magnetic resonance are listed with their important physical constants in Table VI. The difference in the population of nuclei between energy states is usually very small, with lower states being occupied by only a few excess nuclei per million. If the states were equally populated, net absorption would equal net emission and no absorption could be observed. Radiationless relaxation processes from high- to low-energy states are therefore of fundamental importance for the maintenance of an absorption signal. Furthermore, such relaxation mechanisms are dependent on the nuclear environment and proximity to other nuclei, so that relaxation rates can provide chemical data.

TABLE VI Common Nuclear Magnetic Resonance Nuclei of Quantum Spin $\frac{1}{2}$

Isotope	NMR frequency (MHz) at 20 kG	Natural abundance (%)	Relative sensitivity per nucleus
^{1}H	85.2	99.98	1.000
^{2}H	90.8	0.02	1.21
^{13}C	21.4	1.11	1.59×10^{-2}
^{15}N	8.6	0.37	1.04×10^{-3}
^{19}F	80.1	100	0.834
^{31}P	34.5	100	6.64×10^{-2}
^{195}Pt	18.3	33.7	9.94×10^{-3}

There are two distinct types of relaxation processes. One is known as spin–lattice or longitudinal relaxation and results from the interaction of absorbing nuclei with the rest of the sample (lattice), which in the standard liquid phase is in vigorous vibrational and rotational motion. Due to the large number of random magnetic components offered by the lattice, there exists a good probability that the phase and frequency of some of the lattice nuclei match the precessional frequency of the high-spin-state nuclei so that energy exchange can take place. The spin–lattice relaxation time T_1 is defined as a measure of the average lifetime of nuclei in the higher energy state. This lifetime is a function of the lattice mobility and can be used to probe microstructures and viscosities. The availability of a paramagnetic molecule or ion in the lattice provides a strong fluctuating magnetic field source, which can substantially shorten T_1. The second relaxation process is known as spin–spin or transverse relaxation and is denoted as a value T_2. This value is actually a sum of a number of distinct processes that increase absorption bandwidth by decreasing T_2. Some of these processes include energy transfer between closely spaced-identical nuclei that exist in high- and low-energy states, so that the lifetime in any one state is shortened, and also the effects of other magnetic nuclei whose spins create local magnetic fields, which align with or against the permanent externally applied field so that a range of absorption frequencies exist.

Chemical information can also be derived from the position of absorption signals measured on an energy basis and the fine structure of such absorption signals. The position of absorption bands is referred to as the chemical shift and is measured relative to a standard substance. The effect is

due to electrons that circulate in the molecule containing the absorbing nucleus. Circulation of charge creates a magnetic field, which can reduce or enhance the applied field in a local area so that

$$H_0 = H_A(1 - \sigma),$$

where H_0 is the resultant field and is equal to the original applied field H_A corrected for the shielding parameter σ. A compound is chosen as a standard for calibration if its shielding value is larger than the values commonly associated with the nuclei of interest (e.g., tetramethylsilane for proton magnetic resonance studies). Correlation of the chemical shift with structure can provide useful group identification. Chemical shift values for protons are the most common and are often referred to on a δ or τ scale,

$$\delta \cong \frac{H_{\text{ref}} - H_{\text{sample}} \times 10^6}{H_{\text{ref}}}$$

$$\tau = 10 - \delta,$$

where H_{ref} and H_{sample} are the field strengths required to produce tetramethylsilane and sample resonance, respectively. A summary of some common proton chemical shift values is given in Table VII. Additional information can be garnered from the fine structure of absorption bands, which is known as spin–spin splitting. This occurs when the field about one nucleus is affected by the fields from neighboring nuclei attached to an adjacent atom. The degree of splitting reported as frequency differences and the relative areas under each separate absorption signal can, therefore, provide quantitative information about the chemical environment.

Instrumentation incorporates a high-strength, high-quality magnet, which may be permanent, electrically induced, or superconducting. As the field strength becomes higher, the resolution of the instrument improves, and 600-MHz instruments for proton studies are now available. A pair of secondary field coils are located parallel to the high-power magnet faces for the adjustment of field strength over small ranges. A radiofrequency source consisting of oscillator coils is mounted perpendicular to the magnetic field direction and provides plane-polarized radiation. The signal produced by nuclei in resonance is detected by another coil surrounding the sample and mounted perpendicularly to the radiofrequency source coil. Samples are usually in liquid form and are placed in a narrow glass tube, which is rapidly spun to eliminate the effects of field inhomogeneities. Analysis of solid samples is possible when a special arrangement is available to orient and rapidly spin the sample (>2 kHz) at a "magic angle" in relation to the magnetic field. This angle of 54.7° is dictated by geometrical constraints. The use of Fourier transform methods (see Section III.A.3) has provided a means of amplifying the nuclear magnetic resonance signal so that many insensitive nuclei are now routinely investigated, as shown in Table VI. Samples are irradiated with a broad radiofrequency spectrum for a short period of time. After this pulse of energy is applied, the excited nuclei relax to the lower energy states, providing a time-based free induction decay spectrum. This spectrum represents the overlap of the different resonant frequencies, producing a characteristic envelope of time-dependent oscillations. The time-domain spectrum can be collected in seconds, allowing experiment replication to occur hundreds of times in a practical time period. These spectra can be collected and averaged by computers to provide a tremendous signal-to-noise enhancement and are finally displayed as conventional frequency-domain spectra. Modern techniques now allow two-dimensional analysis where excitation scans across different frequency ranges are concurrently analyzed to provide information about coupling between nuclei.

TABLE VII Correlation of Common Proton Chemical Shifts

Group	δ Scale	τ Scale
Tetramethylsilane	0	10.00
H_3C—C (saturated)	1.3–0.7	8.7–9.3
—CH_2— (saturated)	1.5–1.2	8.5–8.8
H_3C—C=C	1.9–1.6	8.1–8.4
—H_2C—C=C	2.3–1.8	7.7–8.2
H_3C—Ar	2.5–2.1	7.5–7.9
H—C≡C— (nonconjugated)	2.7–2.4	7.3–7.6
H—C≡C— (conjugated)	3.1–2.8	6.9–7.2
H_3C—O—	4.0–3.3	6.0–6.7
—C=CH— (cyclic)	5.7–5.2	4.3–4.8
ArH (benzenoid)	8.0–6.6	2.0–3.4
R—CHO	9.8–9.5	0.2–0.5
R—COOH	11.5–11.0	−1.5 to −1.0

Nuclear γ-ray resonance spectroscopy. This technique is based on the resonance absorption of γ radiation and is more conventionally known as Mössbauer spectroscopy. The source of the radiation is a nuclide fixed in a solid crystal lattice held below the Debye temperature. In this condition, γ radiation of energies less than 150 keV are emitted with no loss of energy. Such quantized γ photons can undergo resonance absorption by the appropriate identical stable nuclide in a solid sample matrix. If the chemical environment of the absorbing nuclide is different from the emitter, energy must be added or subtracted from the radiation to establish resonance. This can be achieved by introducing net motion to the source or absorber to establish a Doppler motion energy term.

Properties of interest for chemical analysis are the so-called isomer shifts, which represent the displacement of resonance from zero Doppler velocity, and two types of line splitting. Splitting can be induced by interaction of the electric field gradient about the nucleus with the electric moment of the excited nucleus and also by interaction of the nuclide magnetic dipole moment with internal or external magnetic fields to produce magnetic hyperfine structure. Instrumentation is based on the γ-ray source, a Doppler motion device, and an energy-proportional γ-ray detector coupled to a multichannel analyzer. Chemical information extracted from Mössbauer spectra indicate oxidation states, sample identity with respect to mineral structure, and quantitative sample mineral compositions. The technique is employed most often for iron and cobalt, though over 50 nuclides have demonstrated the Mössbauer effect.

B. Emission of Electromagnetic Radiation

Matter can exist in a physically excited energy level and can relax to a lower ground state energy by releasing the difference in energy. One mechanism of such relaxation can occur by emission of electromagnetic radiation. Continuous broadband radiation is observed from a thermally excited continuous solid and results from the complicated oscillations originating from the molecules and atoms of the sample. If these molecules or atoms emit independently, then a discontinuous spectrum composed of distinct bands or lines can be observed. The structure of this discontinuous emission is related directly to the quantized electronic energy levels of molecules and atoms as discussed previously, and energy is released in a reversal of the absorption process. Though this can be strictly true for atomic samples, often radiationless decay either eliminates or precedes radiative emission from molecular samples and provides interesting chemical information.

1. Atomic Emission

Emission spectroscopy. Atomic or ionic populations can be prepared by the use of high-energy sources such as flames, electrical discharges, and plasmas. If the energy in these sources is sufficient, electronic excitation occurs by collision processes, and relaxation can take the form of discrete narrow line emission. Since the wavelengths of such emissions reflect the energy differences in the quantized electronic energy level distributions, the emission wavelengths are characteristic of the excited element and can be used for identification purposes. The intensity of emitted lines can be used for quantitative analysis by comparison with calibration standard signals recorded on film or collected by photoelectric detectors. This technique allows a multitude of species to be determined concurrently since characteristic lines originate from all suitable species simultaneously. A schematic representation of a simple instrument would be equivalent to that shown in Fig. 5 if the hollow cathode lamp and modulation system were removed. A more energetic atom source such as an inductively coupled plasma (a zone of highly energetic ionized inert gas such as argon, which provides a high constant operating temperature, a long sample residence time, and a nonoxidating amosphere) is usually employed. Detection limits are generally in the range observed for atomic absorption methods, though the two techniques are complementary, each surpassing the other in the analysis of certain species.

Atomic fluorescence spectroscopy. Fluorescence refers to a process whereby absorption and reemission of radiation are separated temporally. A pulsed source of high intensity such as a laser, electrodeless discharge lamp, gaseous discharge lamp, or specially adapted hollow cathode lamp at the required resonant frequency is used to irradiate an atomic population created usually by a nonflame method. Emission can occur in any direction and is commonly observed by standard atomic absorption instrumentation set at an angle of 90° to the high-intensity source beam. This arrangement has been shown to improve detection limits over those for atomic absorption or emission for up to 10 elements.

X-ray fluorescence spectroscopy. This technique is one of the most widely used for qualitative and quantitative elemental determination for elements of atomic number greater than 8. A beam of X-rays is directed from a source such as a Coolidge tube or radioactive substance to a sample. Figure 4 indicates that such radiation is sufficient for ionization of the core electronic level in atoms of the sample. The process of absorption of X-rays can be used to provide structural information about atomic positions in three-dimensional space by a technique known as Extended X-ray Absorption Fine Structure (EXAFS). Relaxation of such excited elements occurs by the shifting of an electron from a higher energy level to fill the electronic vacancy. The difference in energy between the two electronic levels is released as quantized fluorescence in the X-ray region of the electromagnetic spectrum. A large number of different X-ray energies can be emitted from a single sample simultaneously, necessitating the employment of wavelength- and/or energy-dispersive devices in many cases. Wavelength dispersion is accomplished in a manner analogous to that for grating monochromators, but in this case a collimated beam of fluorescent X-rays is allowed to impinge on a solid crystal of well-defined lattice spacing d (e.g., lithium fluoride, sodium chloride,

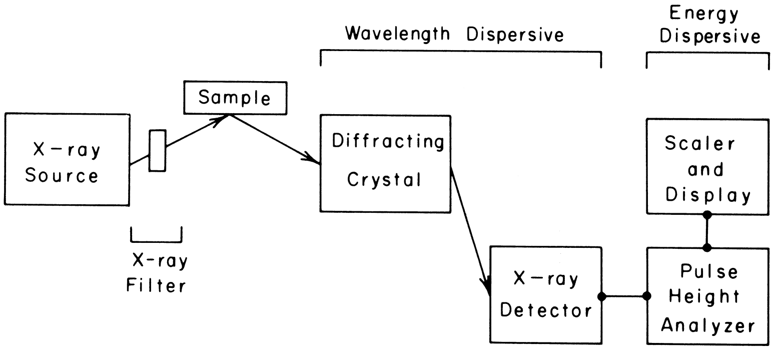


FIGURE 10 Schematic representation of an X-ray fluorescence spectrometer employing both wavelength dispersion and energy dispersion.

topaz) mounted on a goniometer to satisfy Bragg's law,

$$n\lambda = 2d \sin \theta,$$

where n is an integer, λ is the wavelength, and θ is the angle at which constructive interference occurs. Energy-dispersive systems (see Section II.B.3.b) employing pulse height analysis can be used in conjunction with wavelength dispersion to distinguish between two or more energies of photons appearing at one angle due to different values of n, or a multichannel analyzer for discrete energy measurement can be independently employed. A schematic representation of a combined wavelength- and energy-dispersive system is shown in Fig. 10.

2. Molecular Emission

Fluorescence spectroscopy. Molecular absorption of ultraviolet or visible radiation results from a resonant effect between an incoming photon and the electrons of a chromophore, elevating the valence electronic configuration to a higher energy state. Fluorescence is said to occur when the system relaxes to a lower energy state by the release of radiation within 10^{-8} sec after absorption in a process that ceases in less than 10^{-6} sec. Resonance fluorescence occurs when the absorbed radiation is reemitted at the same wavelength. Polyatomic molecules often reemit at wavelengths longer than those absorbed, producing a spectral feature known as a Stokes shift. This is due to nonradiative energy losses by molecular vibration in the electronically excited state. The closely related processes of fluorescence and phosphorescence are summarized in Fig. 11. Phosphorescence occurs from a forbidden excited triplet state after inter-system crossing and may take a period of seconds. This phenomenon is often of very weak intensity and is not as common as fluorescence, thereby limiting its analytical potential. Competition between nonradiative and fluorescent energy loss occurs, reducing the

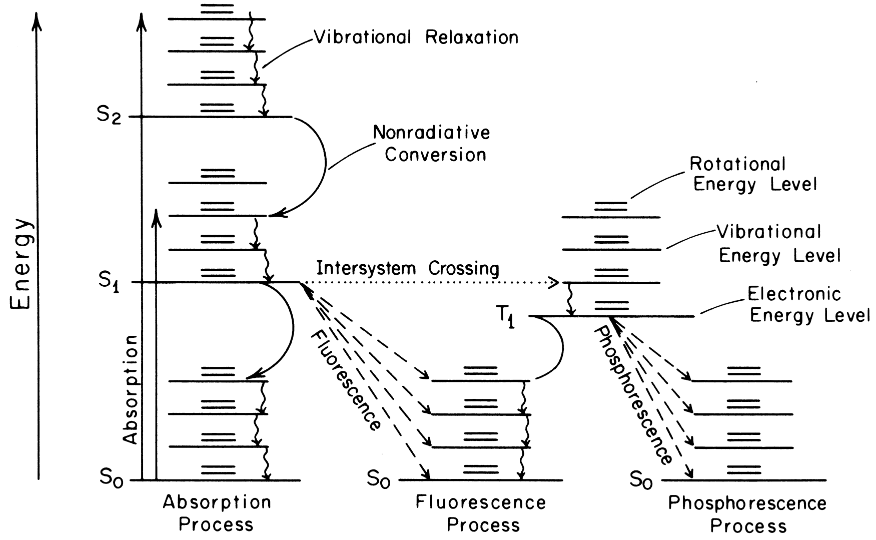


FIGURE 11 Energies of fluorescence and phosphorescence processes.

quantum yield, which represents the ratio of energy released by radiative processes compared with total energy absorption. Ideally, the quantum yield would have a value of unity, but molecular interactions such as intermolecular collisions or intramolecular rotations in the sample matrix can reduce this dramatically. The fluorescent intensity F is therefore a function of the chemical structure of chromophores and molecular interactions and, for dilute solutions with $A < 0.05$, can be written as

$$F = 2.3KAP_0,$$

where K is dependent on the quantum yield, A is absorbance, and P_0 is the incident radiation power. The instrumental components of a spectrofluorimeter are very similar to those for the molecular absorbance technique described previously. Observation of fluorescence occurs at 90° to the angle of source illumination, and an emission monochromator may precede the detector. These instruments are inherently more sensitive than those based on absorbance since at low sample concentrations it is much easier to amplify electronically a small radiant signal superimposed on a dark background than measure a small difference in intensity due to absorption measured on a bright background. These systems can have one to four orders of magnitude better sensitivities than corresponding absorption techniques, though limitations of applicability occur since many chromophores do not undergo fluorescent relaxation. For absorption wavelengths of 250 nm or greater, fluorimetry is the method of choice for direct molecular quantitation at low concentrations. The technique has found great application in organic chemistry, natural product chemistry, and fluorimetric reagent analysis.

Raman spectroscopy. The Raman effect is based on the molecular scattering of incident electromagnetic radiation in the ultraviolet or visible wavelength range. The scattered radiation is quantized due to the interaction with molecular vibrational levels, and therefore, the difference between the incident beam energy and scattered radiation falls within the infrared range. These differences in wavelength are dependent on molecular polarization, as opposed to the dipole moment changes associated with infrared absorption spectroscopy, making the two techniques complementary for structure determination. Spectral analysis shows vibrational bands at values both above and below the wavelength of the monochromatic radiation used for sample irradiation as governed by the physical processes shown in Fig. 12. Energy is absorbed to promote a molecule to a quasi-excited state characterized by a distorted polarized species. Relaxation occurs after 10^{-15} to 10^{-14} sec to the ground state or one of the higher vibrational levels. Spectra are usually plotted as wave number

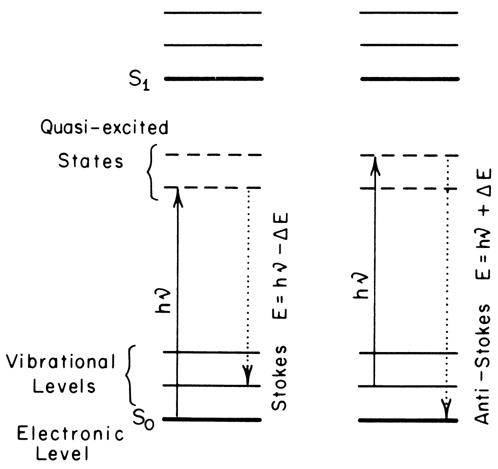


FIGURE 12 Energetics of the Raman process.

shifts from the source line and include both Stokes and the generally weaker anti-Stokes lines. These lines represent infrared absorption as well as active and inactive transitions and can be correlated with the latter technique to distinguish between vibrational frequencies related to dipole moments and induced distortions of electrons from their usual locations in bonds.

The instrumentation used for modern Raman spectroscopy consists of an intense source of optical or ultraviolet radiation, a system for sample illumination, and a monochromator-based spectrophotometric system using standard photomultiplier detection. The incident light path is usually perpendicular to the direction of detection, as in the spectrofluorimetric instruments previously described. Resolution to 0.2 cm^{-1} is possible with the best instruments, and the Raman signal is directly proportional to concentration of the active species. Signal intensity in the absence of absorption is proportional to the fourth power of the source frequency, though wavelengths shorter than 450 nm are not usually employed because of rapid sample decomposition caused by intense energetic radiation. A large variety of surfaces can be studied by use of Surface-Enhanced Raman Spectroscopy (SERS). Local electromagnetic effects on certain roughened conductive surfaces (e.g., Ag, Au, Cu) enhance the Raman signal by factors up to 10^6 by increasing the incident power at the surface and increasing the Raman scattering cross section, making detection and spectral characterization of even submonolayer coverage straightforward.

3. Nuclear Emission

Activation analysis. This technique is based on the measurement of the radioactivity of a sample after activity has been induced by irradiation with nuclear particles, such as thermal neutrons, or highly energetic γ- or X-rays.

Four major distinctions can be made between methods of activation analysis. They include the type of radiation or particle bombardment used for sample excitation, the type of radiation observed in the final emission process, sample preparation by isolation of the species of interest, or direct sample observation without separation by an energy-discriminating device such as a γ-ray spectrometer.

Neutron activation analysis is the most common technique employed today, and the activity A induced in a sample can be related to neutron flux ϕ and sample half-life $t_{1/2}$ as follows:

$$A = N\sigma\phi[1 - \exp(-0.693t/t_{1/2})],$$

where N is the number of sample nuclei, σ is their neutron capture cross section, and t is the time of sample irradiation. From this equation, it can be seen that, after a particular irradiation time, the activity reaches a saturation level where the rate of formation and decay of the active species becomes equivalent. At any one neutron flux, the saturation is determined by the characteristic half-life of the sample, which consequently establishes the analysis time required to attain maximum sensitivity.

The most significant analytical feature offered by activation analysis is its inherent sensitivity, which is suitable for trace analysis to values as small as 10^{-12} g for some species. However, such sensitivities are attainable for only some elements, whereas others may require a total mass of 10^{-4} g before being detectable.

γ-Ray spectroscopy. The ability to measure photon energy accurately at this part of the electromagnetic spectrum is used to great advantage in the identification, discrimination, and quantification of both natural or induced radioactive species. The principle of operation can be considered as proceeding in three phases. An energetic photon interacts with a scintillation crystal–photocathode arrangement or a semiconductor material as the first step in transduction to an electric signal. This electric signal is amplified by conventional electronics so that the output signal energy is proportional to that of the original photon. The second phase involves analysis of the energy of the output signal so that it can be sorted with respect to a preassigned energy resolution into a defined energy level, or "channel." The last phase involves counting and display of the signals in any one channel for quantitative analysis. Both multichannel and single-channel instruments exist and are selected according to cost and flexibility. Multichannel systems can monitor hundreds of discrete wavelengths simultaneously and are commonly used for multielement analysis in conjunction with techniques such as neutron activation analysis.

C. Separation Science

Separation science is a very broad term covering all techniques that employ chemical and/or physical methods to isolate one or more species from some mixture. These methods include solvent extraction, distillation, selective precipitation, crystallization, and chromatography.

1. Principles of Chromatography

Chromatography is a very specialized, yet widely applicable technique of separation science by which one can effectively separate chemically similar substances in complex mixtures. There are three distinct physical methods for attaining separation, yet all are characterized by their common application of a stationary and a mobile phase. The stationary phase provides the chemical surface, which interacts with the components of the mixture to be separated. The difference in the chemistry of these components results in various degrees of molecular bonding or entrapment to the stationary phase. If the mixture is carried across a bed of stationary phase by some mobile phase, the molecules experiencing the least interaction with the stationary phase will tend to advance more quickly than those retained on the bed by stronger molecular interactions. This results in the separation of similar species as influenced by differences in migration rates across the stationary bed. The three distinct separation strategies are summarized below and all assume a common stationary phase.

1. Frontal chromatography. The mobile phase is the sample mixture, and as the sample migrates down the separation bed, the least retained species are concentrated near the leading edge.
2. Displacement chromatography. A sample mixture is applied to one end of a stationary phase, and then a mobile phase is introduced. The mobile phase selectively binds to the stationary phase competing for the same binding sites as the sample. Separation is established as a relative function of the intermolecular interactions of the sample and mobile-phase species with the stationary phase.
3. Elution chromatography. A sample mixture is applied to one end of a stationary phase, which may already contain mobile phase. The mobile phase does not interact significantly with the stationary phase and only carries the sample mixture through the stationary bed. The differences in selective binding of sample mixture components with the stationary phase retards migration of some species so that separation occurs. This is the most analytically useful form of chromatography due to its tremendous power of separation of very similar species.

Stationary phases can be solid surfaces or surfaces coated with liquids, while mobile phases can be liquids or gases. Standard forms that stationary phases take are often very fine granular powders spread evenly on planar glass or polymer supports (e.g., thin-layer chromatography) or packed in columns. Also common are porous polymeric solids usually containing a microcoating of water (liquid stationary phase) or capillary tubes whose inside walls are coated with a liquid stationary phase.

The theory of chromatography has been reasonably well established with a "kinetic" or "rate" theory that describes the broadening of the bands or zones of separated components on a stationary-phase bed, describes their time of appearance at any particular point, and provides details of the separation power or resolution of the particular system employed. Numerous equations describing zone broadening have been proposed, the simplest being the general form of the van Deemter equation as derived for gas–liquid chromatography,

$$H = A + B/V + Cv,$$

where A, B, C represent three kinetically controlled processes known as eddy diffusion, longitudinal diffusion, and nonequilibrium mass transfer, respectively. The term v represents the mobile-phase flow rate, and H is defined as the height equivalent of a theoretical plate. This terminology derives from the plate theory of distillations, where a distillation tower can be divided into a number of equilibria steps or plates. In the chromatographic sense, the height equivalent of such a step represents simply the length L of the separation bed divided by the number of steps in the separation bed N (determined from efficiency of separation):

$$H = L/N,$$

The three kinetically controlled factors responsible for zone broadening can be readily analyzed to improve separation efficiency, which is equivalent to minimizing the value of H. Eddy diffusion refers to the torturous path that sample species must take through a packed particle stationary-phase bed. The distance traveled by different molecules as they migrate across the bed must vary since different paths are highly probable. This results in zone broadening and is critically dependent on average particle diameter and size distribution. Longitudinal broadening refers to the natural diffusion of molecules from a concentrated band toward areas of lower concentration on either side. The most important parameters are those that control molecular diffusion rates in the stationary phase and in the mobile phase. Nonequilibrium mass transfer recognizes the fact the mobile phase may move too quickly to allow concentration equilibrium to be established between the mobile and stationary phases. Factors that control the rate at which equilibrium is attained include the volume and average thickness of the liquid stationary phase and diffusion coefficients and relative equilibrium concentration ratios of the mobile and stationary phases.

A feature of fundamental interest for any chromatographic work is the resolution or separation capability of the technique. One equation useful for quantitative measurement of resolution R can be written

$$R = 2\Delta Z/(W_1 + W_2),$$

where W_1 and W_2 represent the base width of two different peaks in a separation and ΔZ represents the difference in elution of the maxima of the two peaks.

2. Gas Chromatography

This technique is one of the cornerstones of chromatographic analysis, being suitable for the separation of very complex organic and inorganic gas samples. The mobile phase is a carrier gas chosen to maximize separation efficiency as per the van Deemter equation, while concurrently maximizing the sensitivity of the detector. The stationary phase is either a solid or a viscous organic liquid coated on a solid support (hundreds available) or an open tubular capillary column (for very high efficiency separations).

A schematic representation of a gas chromatograph is shown in Fig. 13, and it indicates that the sample is first vaporized by injection into a heated port, then passed through a heated separation column, and finally detected. The small sample size that can be passed by capillary columns usually necessitates supplementation of specialized splitters at the injection port. A splitter accurately and reproducibly reduces the volume of sample that enters the column from standard sample injections done by syringe. Numerous detection devices exist, and the most common are summarized in Table VIII. Separation times of various species eluting from the column are often controlled by varying column temperature by a technique known as temperature programming.

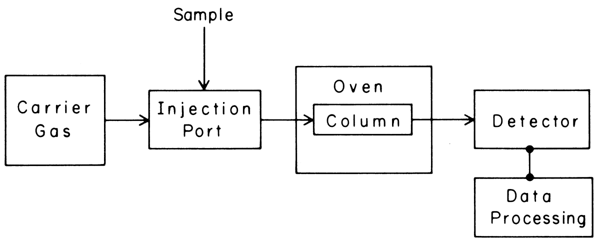


FIGURE 13 Schematic representation of a single-column gas chromatograph.

TABLE VIII Summary of Common Gas Chromatographic Detector Characteristics

Detector type	Selectivity	Limit of detection ($g\ sec^{-1}$)
Thermal conductivity	None	10^{-9}
Flame ionization	Combustibles	10^{-12}
Flame photometric	Sulfur, phosphorus species	10^{-12}
Electron capture	Halogenated species	10^{-13}
Photoionization	None	10^{-14}

TABLE IX Summary of Common High-Performance Liquid Chromatographic Detector Characteristics

Detector type	Selectivity	Limit of detection ($g\ ml^{-1}$)
Refractive index	None	5×10^{-7}
Ultraviolet–visible absorption	Chromophore-containing species	5×10^{-10}
Fluorescence	Fluorophore-containing species	10^{-10}
Amperometric	Electroactive functional groups	10^{-12}
Conductivity	Ionized species	10^{-8}

3. High-Performance Liquid Chromatography

This form of chromatography is very well suited for separations of organic mixtures and often complements gas chromatography, since many organic species cannot be volatilized readily. Separation efficiency can be similar to that obtained from packed column gas chromatography and is usually achieved on an extremely thin stationary phase coated onto small solid particles of micrometer diameter. The small size and high degree of regularity of the packing material provide great resistance to solution flow, thereby necessitating use of high pressure to force the mobile phase through a column. The technique has greater flexibility than gas chromatography since the mobile phase can be easily changed with dramatic effects on resolution. A schematic of such a chromatographic system is shown in Fig. 14, illustrating the use of fluids as the mobile phase. Solvent programming refers to mobile-phase solvent changes during the course of one chromatographic elution and in some respects is similar to temperature programming. A summary of common detectors is provided in Table IX.

4. Ion Chromatography

Ion chromatography has been known for many years, and ion-exchange resins are perhaps the most commonly used stationary phases for general laboratory glass column separations. However, a tremendous advance in this technology has occurred since the mid-1970s, resulting in the evolution of high-performance ion chromatographs capable of concurrent separation of monovalent and multivalent inorganic and organic ions in periods of minutes. A schematic diagram of such an instrument is shown in Fig. 15. The innovations of this instrument lie in the development of new separation column resins of high efficiency and the development of countercurrent fiber-based suppressor columns that eliminate all but analyte ions from the mobile phase. The elimination of all ions other than the analyte ions is accomplished by use of a semipermeable membrane which can only pass ions of one type of charge. A suppressor solution is maintained on the outside of the membrane and functions by exchanging ions with the mobile phase, maintaining charge balance and resulting in conversion of ions to nonconductive soluble compounds. For example, sodium ions of $2Na^+/CO_3^{2-}$ in the mobile phase can be exchanged with hydronium ions of $2H_3O^+/SO_4^{2-}$ from the suppressor solution, leaving H_2CO_3 as a nonductive compound in the mobile phase. The conductivity detector senses only the presence of analyte ions without a large background signal and can therefore provide detection limits of $10^{-10}M$ concentrations or better.

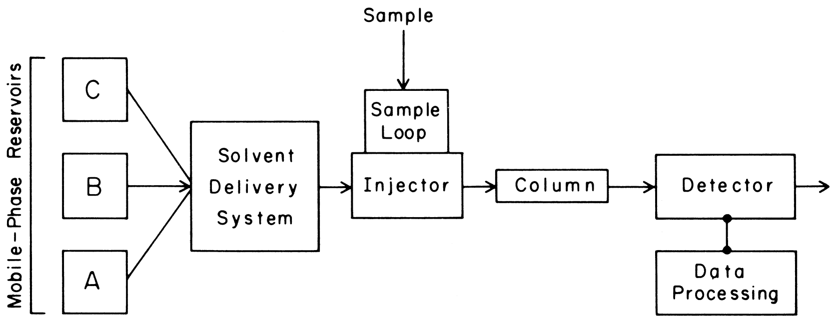


FIGURE 14 Schematic representation of a solvent-programmed, high-performance liquid chromatograph.

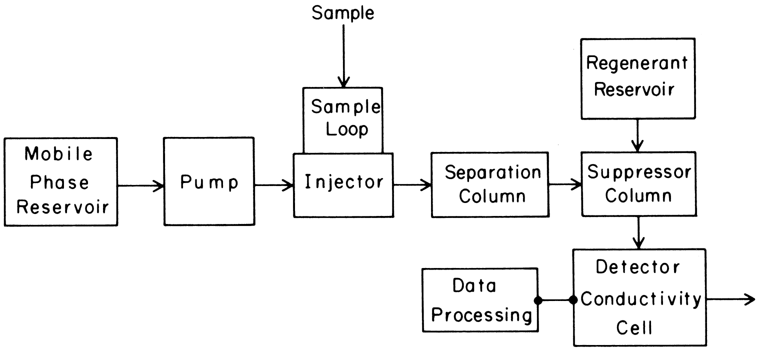


FIGURE 15 Schematic representation of a high-performance ion chromatograph.

5. Supercritical Fluid Chromatography

Supercritical fluids exist at temperatures and pressures above the supercritical point of a compound (e.g., where a gas at STP exists as a fluid). As mobile phases for chromatography, supercritical fluids provide properties that are intermediate to liquid and gas chromatography. The solvent strength is closely related to density and offers another operating parameter in the form of pressure–density programming. The low densities and high diffusivities of gases give gas chromatography better resolution per unit of time than liquid or supercritical chromatography, but gas chromatography is limited to the analysis of low molecular weight, volatile, and thermally stable compounds. High liquid densities provide liquid chromatography with an excellent range of solvating power, but this technique is relatively show and is limited by a lack of sensitive universal detectors. The ability of supercritical fluids to solvate compounds at lower temperatures by controlling the mobile phase density makes it possible to readily analyze thermally labile and nonvolatile compounds. The mobile phase is pumped as a liquid and heated to above the critical temperature before passing through the column as a supercritical fluid via an injection loop valve. A pressure restrictor is incorporated after the column to ensure that conditions throughout the column remain supercritical. The technique uses either packed liquid chromatography columns or open-tubular fused-silica columns, with a variety of nonpolar, polar, or novel chiral-bonded stationary phases.

The most commonly used mobile phase in SFC is carbon dioxide (CO_2) because it has a critical temperature (T_c) of 31.05°C and a critical pressure (P_c) of 72.9 atm, which are easy to work with; it is nontoxic and nonflammable; and it is readily available at low cost. Modifiers such as methanol are sometimes added to change solute elution characteristics.

6. Chromatographic Methods in Biochemical Science

Biochemists address a chemically limited, yet extremely complex area of organic chemistry. The complexity has resulted in the evolution of specialized separation techniques suitable for macromolecules that are difficult to determine by gas and high-performance liquid chromatographic systems. Though many specialized separation methods are employed by biochemists, two techniques have found widespread use in many areas of chemistry.

Gel permeation (or size-exclusion) chromatography is a technique based on molecular separation partially by size rather than chemical interaction with a stationary phase. A water-swollen polymeric gel acts as the stationary phase and is designed by polymeric cross-linking to contain a certain size distribution of pores or cavities. As a sample passes through the polymeric sieve in a water-based mobile phase, smaller molecules diffuse into the pores more readily than larger macromolecules, so that the latter elute first while the progress of the smaller molecules is retarded. The most popular polymer for this type of separation is prepared by cross-linking epichlorohydrin with the polysaccharide dextran.

Electrophoresis is a second specialized technique, employing for analysis the migration of ionic molecules or aggregates in an applied electric field. Commonly, a stationary phase such as a water-swollen gel saturated with an electrolyte is placed between two do high-voltage (hundreds to thousands of volts) electrodes. The sample migrates through the gel as a function of its charge-to-mass ratio, as well as being influenced by all the factors previously described in Section II.C. Staining of the gel after the application of the voltage for a defined time allows visualization of sample species. It is possible to arrange the instrumentation to use standard high-performance liquid

chromatographic detectors for sample elution analysis and subsequent species collection.

D. Analytical Electrochemistry

Electroanalytical chemistry includes a broad range of techniques that have as their focus the fact that the analyte participates in a galvanic or electrolytic electrochemical cell. All techniques can be classified into one of three major areas: those that measure electrical properties of the cell, those that measure cell electrical properties as a function of a chemical reaction in the electrolyte, and those that physically collect the analyte at an electrode for further analysis.

1. Conductimetry

When an electrical potential is applied between two electrodes immersed in an electrolyte solution, ions are accelerated and migrate to electrodes of opposite charge. The rate of ion migration is a function of the applied potential acting as a driving force countered by frictional forces and electrophoretic and relaxation effects in solution. Conductance is represented as reciprocal resistance and can be corrected for the cross-sectional area and length of the conduction solution (specific conductance), or it can be stated as the specific conductance of 1 equivalent of solute (equivalent conductance). A linear relationship exists between equivalent conductance and the square root of the concentration of a strong electrolyte. The measurement of solution conductance is highly sensitive, though not selective for any particular species. The device finds application as a detector for ion chromatographic instruments, due to its inherent sensitivity. Conductance can be employed as a parameter to follow certain titrations such as those involving neutralization, precipitation, and complexation reactions that consume or generate ionic species.

The conductance experiment uses an ac source to avoid faradaic currents, with oscillation frequencies from 60 to 1000 Hz. A Wheatstone bridge assembly can be used as the detection element if the electrochemical cell acts as one arm of the bridge, though other electronic assemblies are suitable for direct instrumental measurement of conductance. The electrochemical cell consists of two large area platinized platinum electrodes placed in fixed orientation with defined separation.

2. Potentiometry

The measurement of the electromotive force between two electrodes can provide information about the activity or concentration of an ion in the analyte solution. The potential of any electrode (E) can be related to the activities of the oxidized and reduced forms of any species (a_{ox}, a_{red}) by the general form of the Nernst equation,

$$E = E^\circ - \frac{RT}{nF} \ln \frac{a_{\text{red}}}{a_{\text{ox}}},$$

where E° is the standard electrode potential, R is the molar gas constant, T is the absolute temperature, n is the charge associated with the redox reaction, and F is the faraday.

Indicator and reference electrodes. Complete analytical electrochemical cells are composed of combinations of an indicator and a reference electrode. Each of these electrodes contributes half-reaction chemistry; in combination they provide a complete redox reaction. The indicator electrode is the analyte activity sensing element and develops a potential E_{Ind} relative to the reference electrode, which represents a constant known potential E_{Ref} independent of the composition of the sample solution. The measured cell potential E_{cell} can be written

$$E_{\text{cell}} = E_{\text{Ind}} + E_{\text{Ref}} + E_{\text{Junc}},$$

where E_{Junc} represents the liquid junction potential originating at the interfaces of the electrodes in the sample solution due to the unequal distribution of ions as a function of their different rates of migration across the interface. Ideally, the reference electrode potential is constant, and the junction potential is constant or negligible. A number of reference electrodes are available, the most common being silver/silver chloride, calomel, or hydrogen gas systems.

Indicator electrodes are generally employed for titration analysis where the cell potential varies as ions are consumed or produced in a chemical reaction. At least four important types of indicator electrode can be identified. The simplest systems involve a strip of material suitable for the provision or acceptance of electrons, such as gold, platinum, or carbon, immersed in a solution containing both the oxidized and reduced forms of some species so that a redox potential can be measured. Another situation involves the use of a metal that is reversible with respect to analyte ions of the metal so that the metal electrode participates in mass and charge balance rather than only charge balance. A third class of electrodes uses two interfaces, as in the case of the silver/silver chloride reference system, where the metal is coated with a sparingly soluble salt and is suitable for titration analysis as well as determination of the activity of the metal or the anion in the salt coating. Another class of electrodes representing a specialized series of membrane-based devices designed to monitor the activities of certain ions selectively is described in the following two sections.

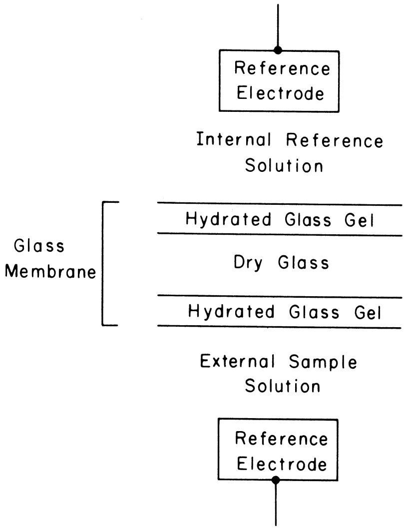


FIGURE 16 Construction and hydration properties of an ion-selective glass electrode.

Glass electrodes. Glass electrodes are ion-selective electrodes based on the chemical properties of a glass membrane of defined chemical composition. Alteration of the glass chemistry to contain variable quantities of Na_2O, CaO, SiO_2, and Al_2O_3 produces chemically active binding sites in the glass that have hydronium, sodium, or simple cation selectivity. The construction of a typical electrochemical cell based on a glass electrode is shown in Fig. 16, including a representation of the physical properties of the sensing membrane. The internal solution contains a fixed concentration of the cation of interest and therefore fixes the internal electrode surface potential, while that in the external solution varies. The electric potential develops only in each hydrated gel layer based on an ion-exchange principle that leads to a phase boundary potential. The entire glass membrane is nominally 50–100 μm in thickness, but the hydrated glass accounts for little of this, being only 5–100 nm thick. Conduction within the dry glass is due to the cation of lowest charge and is not related to penetration by a substantial amount of cations.

Ion-selective electrodes. The glass electrode is one example of an ion-selective electrode; three other specialized types exist as defined by their membrane construction. A summary of these is given in Table X. All these devices are governed by modifications of the Nernst equation, as shown for the fluoride electrode at 25°C,

$$E_{\text{cell}} = \text{const} - \frac{0.059}{n}\log\left(a_1 + K_{1,2}a_2^{n/z}\right),$$

where the cell potential E_{cell} is related to a constant including internal and external reference electrode potentials, the activity a_1 of the analyte, and the activity a_2 of an interfering ion of charge Z adjusted for response with a selectivity factor $K_{1,2}$ representing the ratio of selectivity for ion 1 over ion 2. All selective electrodes suffer from interferences caused by competing species, and these must be recognized and controlled in any quantitative assay.

It is possible to produce hybrid potentiometric electrodes suitable for the detection and measurement of certain gases and organic molecules. An ion-selective electrode can be used as a transducer to measure changes in ion activity caused by enzyme–substrate reactions or dissolution of gases at small aqueous volumes trapped at the selective electrode surface. For example, an enzyme trapped in a hydrophilic water-swollen gel can be attached to the surface of a glass electrode. When placed in a substrate-containing sample solution, the enzyme–substrate reaction may produce a certain amount of hydronium ion product, which can be quantitatively detected as a local pH alteration by the glass electrode. Further extension of this technology makes use of enzyme-linked immunoassay. An antibody that is attached to an enzyme provides immunochemical selectivity in a competitive binding or sandwich assay strategy, while the enzyme provides the electrochemical signal.

3. Voltammetry and Polarography

Voltammetry refers to a broad range of techniques that compare current–voltage relationships between a working and a reference electrode. The best known of these techniques is polarography, which is based on an electrolytic reduction or oxidation at a microelectrode when the rate of the redox reaction is controlled by analyte diffusion to the working electrode surface. Standard working electrodes are easily polarized microelectrodes such as the mercury drop, rotating platinum, glassy carbon,

TABLE X Common Ion-Selective Electrodes

Membrane type	Construction	Typical ion selectivity
Glass	Thin, specially formulated glass	H^+, monovalent Group I cations
Solid state	Pellets or polymer matrix of insoluble salts	Cl^-, Br^-, I^-, Cu^{2+}, Cd^{2+}, Pb^{2+}
	Single crystal of LaF_3 doped with EuF_2	F^-
Liquid ion	Hydrophobic charged ion-exchange liquid trapped in porous lipophilic diaphragm	Ca^{2+}, Mg^{2+}, NO_3^-, ClO_4^-, Cl^-
Neutral carrier	Neutral chelating agent in liquid trapped in porous diaphragm or supported in polymer matrix	K^+, Na^+, Rb^+

and wax-impregnated graphite electrodes. Conventional electrochemical cells are designed on the basis of three-electrode (working, reference, counter) potentiostatic control, which allows work in solvents of high resistance.

Polarography applies a continuous or pulsed or oscillating potential ramp to the working electrode and observes the generation of a redox current as a function of applied potential. Three mass transfer processes can control the rate of analyte deposition at the surface of the working electrode: convection in the sample solution, migration of charged ions in the applied electrostatic field, and diffusion due to a concentration gradient. Experiments are usually conducted without stirring and in 100-fold concentration excess of an electrochemically inactive electrolyte to eliminate the first two processes. When the working electrode potential reaches a threshold value, dependent on the standard potential for the half-reaction of interest, oxidation or reduction of the analyte at the electrode surface begins. Concentration diffusion begins from bulk solution and eventually reaches a steady state, indicated experimentally as increased but constant redox current (Fig. 17). The potential at the midpoint of the polarographic step is known as the half-wave potential $E_{1/2}$ and is related to the applied potential E_{appl} for the half-reaction, the measured current i, and the diffusion-controlled current i_{d} as

$$E_{\text{appl}} = E_{1/2} - \frac{RT}{nF}\log\frac{i}{i_{\text{d}} - i},$$

The current, which is diffusion limited, is a function of electrode geometry as well as other parameters, as shown in the simple form of the Ilkovic equation for a classical dc polarographic experiment using a dropping mercury electrode,

$$i_{\text{d}} = 607nCD^{1/2}m^{2/3}t^{1/6},$$

where i_{d} is the average diffusion current in microamperes for each drop, n is the redox electron transfer, D is the analyte diffusion coefficient in square centimeters per second, C is the analyte concentration in millimoles per liter, m is the rate of mercury mass flow in milligrams per second, and t is the drop time in seconds. This equation demonstrates that the diffusion-limited current is directly proportional to analyte concentration.

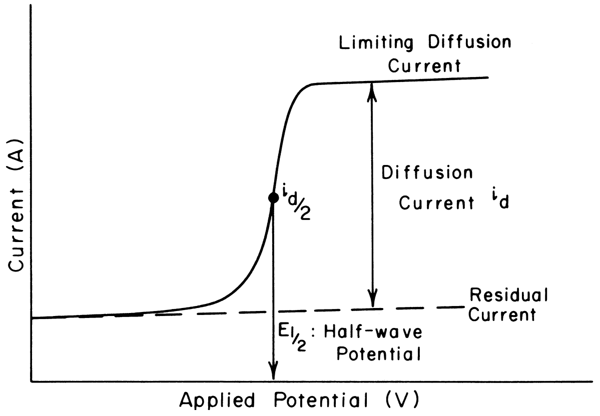


FIGURE 17 Electrochemical profile observed in a classic polarographic experiment.

TABLE XI Common Techniques for Polarographic Analysis

Technique	Detection limit (*M*)	Resolution (mV)
Classical polarography	10^{-5}	200
Rapid-scan polarography	10^{-7}	50
Cyclic voltammetry	10^{-7}	50
Differential pulsed polarography	10^{-8}	50
Square wave voltammetry	10^{-8}	50
Stripping voltammetry	10^{-10}	100

Numerous procedures for applying electric potential have been developed to increase the sensitivity of polarographic methods. The most sensitive of these procedures makes use of application of short (millisecond) voltage pulses and sample the current which flows as a function of the pulse regime. These procedures concentrate on maximizing the desired faradaic current while minimizing background signals originating from nonfaradaic processes. A summary of these techniques is presented in Table XI to indicate electric potential application methods. The speed of techniques such as square wave voltammetry makes such methods preferred for many analytical situations. Polarographic techniques are often used in analyses of trace concentrations of toxic metals and are applicable to almost every element in the periodic table. Analyses of concentration and reaction mechanisms of organic species containing certain functional groups are also important.

4. Amperometric Titrations

The methods of polarography can be used to follow the progress of a titration that produces or consumes some electroactive species. The voltage applied to the working electrode is constant and sufficient to provide the diffusion-limited current for the observable analyte. The current is then recorded as a function of the volume of titrant added to the sample solution. Since the working electrode is a microelectrode with an area of only a few square millimeters, and the current that flows is usually of the order of microamperes or less, only an insignificant amount of analyte ion is removed from the sample solution. This implies that the redox process involved at the working electrode has a negligible effect on the chemical titration reaction and does not introduce significant error in most instances.

5. Coulometry and Electrogravimetry

These techniques are dependent on an electrolysis that attempts to oxidize or reduce completely all of an electroactive analyte species in a volume of sample solution. These methods have the distinct advantage of requiring no calibration since they are dependent on total electron transfer, which can be quantitatively related to the total mass of the analyte. Coulometry refers to the integration of electric current with time as electrolysis proceeds either at a constant current or at a constant voltage. The latter parameters are chosen to optimize analysis time and limit interferences. For constant current, the number of coulombs of electrons consumed or produced, Q, is related to the current I and time t by

$$Q = It.$$

For variable current, the relationship becomes

$$Q = \int I\,dt.$$

Typical electrochemical cell arrangements employ a three-electrode potentiostat with a platinum working electrode. It is possible in many cases to arrange the experiment so that the electrolysis deposits the analyte quantitatively on the working electrode. The mass change of the electrode can be related to the number of coulombs of the charge associated with the deposition to determine equivalent weight. These techniques have only moderate sensitivity but provide accuracy and precision to relative values of tenths of a percent.

E. Mass Spectrometry

In mass spectrometry, molecules are both ionized, positively or negatively depending on the imposed conditions, and fragmented. The particles are sorted out into groups according to their mass-to-charge ratio. A plot of the count of the ions against the mass of the different types of ions is called a mass spectrum. This record is valuable analytically in that it is characteristic of each chemical compound. The instrument required for this purpose can be divided into several main components: (1) a device for introducing the sample or inlet system, (2) an ion source, (3) the spectrometer for sorting out ions or ion analyzer, (4) a detector of ions, (5) a recording system, (6) a vacuum system, and (7) a complete control system including computer for data handling and so on. There are a wide variety of such instruments, which can be classified according to the methods of ionization and/or procedures for ion analysis. Here, we shall concentrate on certain areas.

1. Ion Sources

The common features of all ion sources are that they incorporate techniques for producing ions and giving them kinetic energy (acceleration used for introduction into the spectrometer). The electron-impact source is the most widely used (Fig. 18). In this device, molecules in the gas phase, obtained from the inlet system, are subjected to a stream of accelerated electrons (usually at a potential of 70 V), and the resulting collisions between the particles cause molecular ionization and fragmentation. These species are then injected into the spectrometer by the accelerating slits. Another source is based on chemical ionization. Here, the analyte molecule is ionized by an ion–molecule reaction, not directly by electrons as mentioned above. A reagent gas is ionized by electrons, for example, methane,

$$CH_4 + e^- \rightarrow CH_4^+ + 2e^-,$$

with further reactions:

$$CH_4^+ + CH_4 \rightarrow CH_5^+ + CH_3$$
$$CH_3^+ + CH_4 \rightarrow C_2H_5^+ + H_2.$$

These particles can react with the sample molecule (MH) in a number of ways:

$$CH_5^+ + MH \rightarrow CH_4 + MH_2^+$$
$$CH_3^+ + MH \rightarrow CH_4 + M^+$$
$$CH_4^+ + MH \rightarrow CH_4 + MH^+$$

The advantages of this technique are that less energy is transferred to the analyte molecule, causing less fragmentation, and the fragmentation can be controlled by the choice of reagent gas.

The field ionization technique uses the effect produced on molecules by a high electric field (10^7–10^8 V cm^{-1}). In an appropriate electrode design, the electric field is high enough to rip electrons from the molecule of interest.

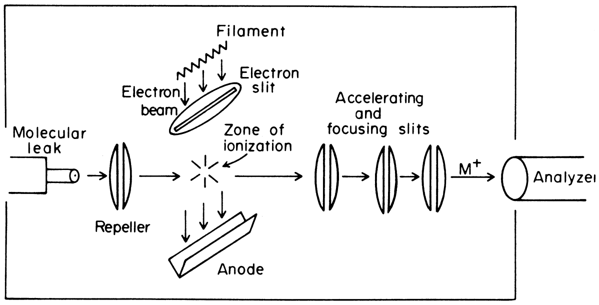


FIGURE 18 Typical electron-impact source for mass spectrometry.

Other sources are field desorption, spark source, thermal ionization, fast atom bombardment, and secondary ion mass spectrometry (see SIMS, Section G.4, Surface Analysis).

2. The Spectrometer

The purpose of the spectrometer is to separate the ions emanating from the ion source as efficiently as possible. Quantitatively, this is expressed as the resolving power of the instrument, which is defined as the ratio $M/\Delta M$, where M and $\Delta M + M$ are the mass numbers of two neighboring peaks of equal intensity in the spectrum. The key factor here is the ability to distinguish M and $M + \Delta M$; usually, this is said to be achieved when the "valley" between the two peaks is no more than 10% of the intensity of M or $M + \Delta M$. Resolution is controlled by a number of instrumental factors, including the method of ion separation.

One widely used system of separation is the magnetic deflection apparatus. Here, a controllable magnetic field causes ions to deflect along curved paths according to their mass-to-charge ratio. The instrument is designed such that only those ions that follow the path coinciding with the arc of the analyzer tube (Fig. 19a) in the magnetic field are brought to a focus close to where the ion detector is located. Thus, the analyzer separates the ions into streams of different M/e,

$$M/e = H^2r^2/2V,$$

where M is the mass of the ion, e is the charge, H is the magnetic field, r is the radius of curvature of the analyzer tube, and V is the accelerating potential employed in the ion source. Usually, V is swept to obtain a mass spectrum. In the time-of-flight mass spectrometer, ions of different mass, which are produced by pulsed electron beam impact in the ion source, are accelerated to the same kinetic energy. The ions are then allowed to drift in space down a tube of particular length before they are detected (Fig. 19b). Because they have different velocity, the transit time t varies as

$$t = L\left[\frac{M}{e}\frac{1}{2V}\right]^{1/2} \mu\text{sec}.$$

This instrument is useful for the study of fast reactions. An additional spectrometer known as the quadrupole device is based on the passage of ions through a region between four, short parallel metal rods of alternating electrical voltage and provides low resolving power. This is widely used because of its simplicity, although its physics is very complex.

All the spectrometers described to this point are called single-focusing systems. These are characterized by low resolving power, which is due to the variety of kinetic energies produced in the ion source. In a double-focusing spectrometer (Fig. 19c), an electrostatic field for ion deflection is introduced between the ion source and the magnetic deflection device. The electrostatic analyzer ions are effectively focused into highly defined kinetic energies before they pass into the magnetic analyzer. The resolving power of such instruments is on the order of 50,000.

Another relatively new device known as the ion cyclotron system, which is not a double-focusing system, can provide resolution to well over 100,000 and can operate in a Fourier transform mode, offering speed and sensitivity.

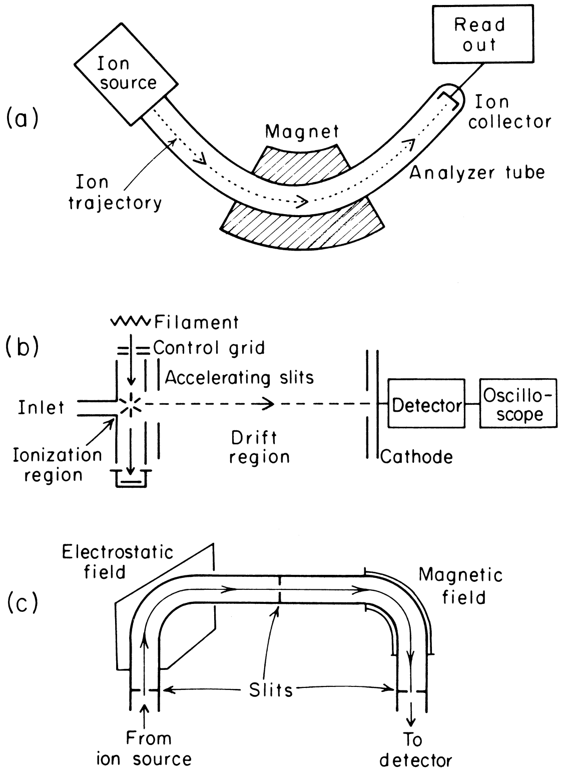


FIGURE 19 Magnetic deflection (a), time-of-flight (b), and double-focusing (c) mass spectrometers.

3. Other Components

The sample inlet system for a typical mass spectrometer is versatile enough to handle gas, liquids, and solids. The device is usually held at ~200°C and 0.02 torr pressure. Accordingly, any solids must have a sufficient vapor pressure under these conditions to allow transport to the ion source as a gas before a spectrum can be recorded. Modern instruments usually incorporate an electron multiplier or channel electron multiplier array as a detecting system. Both these devices work on the principle of electrons released from a material on ion impact. The electrons are

then amplified in number before measurement of the current. It goes without saying that nearly the entire mass spectrometer must be kept under high vacuum with diffusion pumps, backed by rotary oil pumps. If the spectrometer is connected to a gas chromatograph, an interface is inserted between the inlet system and the chromatograph to allow sample transport but to remove carrier gas. Effusion, membrane, and jet separators are used for this purpose. Finally, most modern high-resolution instruments utilize dedicated computers not only for data handling, but also for total control of the instrument. Many facilities are incorporated in these configurations—for example, comparison "libraries" of recorded mass spectra of thousands of known compounds.

4. Analytical Use

The mass spectrometer is clearly of tremendous importance in the identification of chemical compounds. This is because the ionization of compounds produces unique fragmentation patterns. Accordingly, the correlation of mass spectra with molecular structure provides a beautiful foundation for chemical analysis. In identification procedures at a reasonably low level, we can use the mass spectrum in two distinct ways. First, the highest mass peak in the spectrum (Fig. 20) usually is associated with the molecular weight of the compound, a particularly valuable piece of initial information. Furthermore, at masses 1 and 2 larger than this value, we find the effect of naturally occurring heavier isotopes. This aids us in deciding on a formula for a particular molecular weight when we might have several choices. Second, fragmentation produces a set of characteristic mass peaks for each compound. Destruction of a particular ion occurs because there is enough residual energy to cause bond breakage, as well as removal of electrons. Fortunately, a number of rules seem to apply

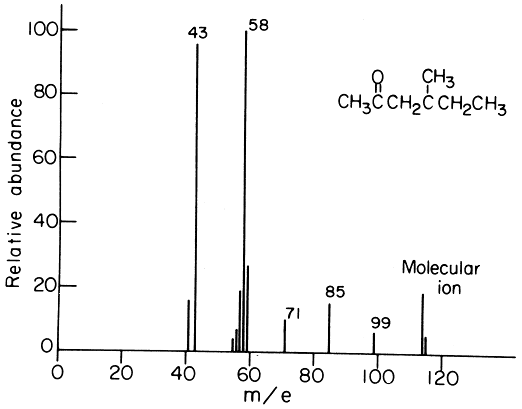


FIGURE 20 Typical mass spectrum of an organic molecule with identified mass peaks.

for bond breakage, which allow us to predict the original structure of a particular species. For example, ring compounds usually contain mass numbers characteristic of the ring (due to its inherent stability). As can be appreciated, the whole philosophy behind working out the structure of a complete unknown from its mass spectrum is akin to the development of a solved jigsaw puzzle. This procedure is quite different from the use of the computer to match the mass spectrum of an "unknown" with its memory bank of data, as mentioned above. In this case we are using the "fingerprint" approach to identification.

F. Thermal Methods

Thermal methods of analysis involve the measurement of a physical property of a sample as a function of controlled temperature. Equipment is available for detecting transition temperatures, energies of transitions, weight loss, size changes, elasticity, and changes in solution temperature plotted as a function of added volume of a reagent.

1. Thermogravimetry

In thermogravimetry (TGA), weight change associated with a transition in the sample as a function of temperature is measured. For example, if water molecules of hydration can be removed thermally, they can be detected by apparatus for TGA. Practically, the sample is placed in a crucible, which is part of an automatic recording analytical balance. The sample configuration is then introduced into a furnace in which the temperature is measured accurately (usually room temperature to ~1200°C). The temperature of the furnace is raised gradually, often in the heating range 5–10°C/min, and any weight losses (increases) are recorded. From plots of loss (increase) in weight vs temperature, information about the gross structure of the sample components and reaction mechanisms can be obtained. The technique is often employed in conjunction with other thermal methods such as differential thermal analysis (DTA) and differential scanning calorimetry (DSC).

2. Differential Thermal Analysis and Differential Scanning Calorimetry

In DTA, the temperature of a sample is compared with that of a reference substance, and the liberation or absorption of energy associated with various sample transitions can be correlated with the difference in the two temperature values. As in TGA, the temperature of the sample is raised in a programmed fashion. The results inform the analyst as to the occurrence of exothermic or endothermic reactions. The instrument itself consists of a furnace

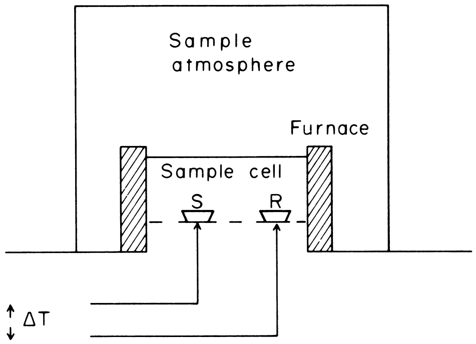


FIGURE 21 Apparatus for differential thermal analysis. S = sample pan; R = reference pan.

that incorporates a heating block for sample and reference material into different chambers (Fig. 21). The block temperature is measured at 5–10°C/min, and the difference in temperature is measured by thermocouples, which are in close contact with the sample and reference material. The overall temperature of the furnace is also recorded. In a rather closely related manner, DSC involves the addition or substraction of thermal energy in order to maintain both sample and reference materials at the same temperature.

The three techniques described above are often correlated with one another in the thermal analysis of a sample, since weight changes are often accompanied by the absorption or release of thermal energy. In DTA or DSC, when an endothermic change occurs, the sample temperature lags behind the reference temperature; the opposite is the case for exothermic processes. Typical TGA and DTA plots for an inorganic material, $Ca(CH_3COO)_2{\cdot}H_2O$, are shown in Fig. 22. In an argon atmosphere, the first endothermic signal matching a weight loss is associated with the removal of water. The second and third of the endothermic peaks, which correlate with weight changes, are caused by the loss of CO to produce $CaCO_3$ and the loss of CO_2 to yield CaO, respectively.

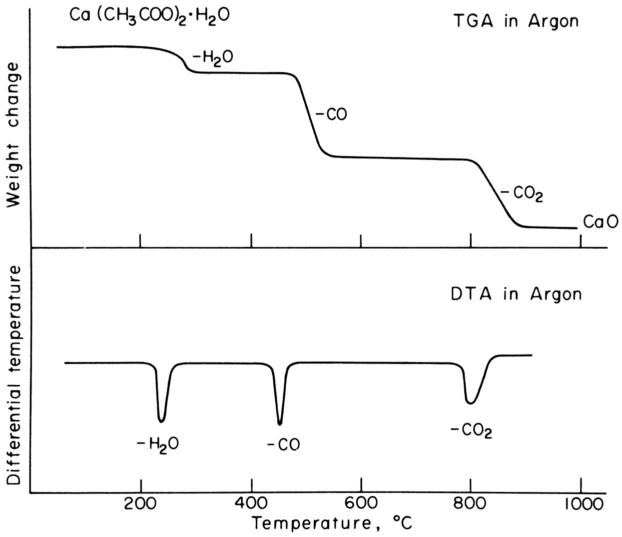


FIGURE 22 Thermogravimetric and differential thermal analysis profiles for $Ca(CH_3COO)_2{\cdot}H_2O$.

Thermal analysis of polymeric materials can be particularly fruitful for the analyst. Melting points, phase transitions, pyrolysis, and curing conditions can all be gleaned not only from the temperature positions in DTA (DSC), but also from the width of endothermic and exothermic peaks. In addition, it is often possible to analyze gases that are liberated from the sample by gas chromatographs and mass spectrometers.

3. Thermometric Titrations

In this technique, a titration of the sample (enclosed in a Dewar container) is performed with an automatic motor-driven buret at the same time as the sample solution temperature measurement. Temperature changes are very small (0.001–0.2°C), and therefore temperature measurements must be made with considerable accuracy. The method yield heats of reaction for a particular chemical system. Applications of the method include determination of the concentration of an unknown, reaction stoichiometry, and thermodynamic parameters. These can be performed in nonaqueous solvents and fused salts.

G. Electron Spectroscopy

When a sample is exposed to electromagnetic radiation of sufficiently short wavelength or an electron beam with particles of sufficient kinetic energy, an emission of electrons can be observed. In the former case we can ascribe electron ejection in terms of a photoionization effect, whereas in the latter we have ionization produced by electron impact. Recent years have seen the development of a family of techniques in which the kinetic energy and number of ejected electrons are measured. Typically, these methods are classified as electron spectroscopic techniques. The overall instrument consists of a sample inlet system, ionization chamber, source of ionizing particles or radiation, kinetic energy analyzer, electron detector, control and recording system, and vacuum configuration (Fig. 23). As

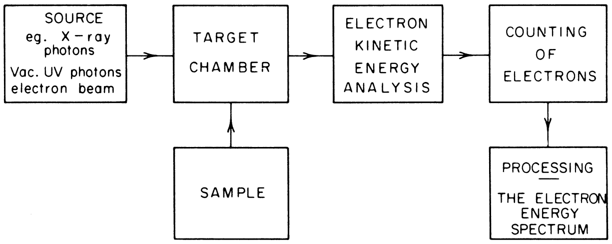


FIGURE 23 General schematic representing the experiment of electron spectroscopy.

with the mass spectrometer, an electrostatic field can be used to focus electrons of certain energy at an exit slit ready for counting (detection). The experimental plot of number of electrons vs their kinetic energy is called an electron energy spectrum. The individual techniques in electron spectroscopy are classified according to either the method of inducing ionization or the nature of the process that accompanies the emission of electrons.

1. Vacuum Ultraviolet Photoelectron Spectroscopy

In vacuum ultraviolet photoelectron spectroscopy (UPS), the sample atom or molecule is exposed to radiation in the vacuum ultraviolet region of the electromagnetic spectrum. A readily available source of radiation is the helium discharge lamp, which produces a sharp HeI line at 21.2 eV. Since the energy required for photoionization of sets of valence electrons is in the vicinity of 6 eV to this energy, we obtain a polyenergetic emission of electrons described by the Einstein relation

$$E_n = h\nu - I_n,$$

where I_n is the ionization energy of the nth species of electron and E_n is the kinetic energy of the bunch of electrons ejected by a photon of energy $h\nu$. Usually, we form an experimental plot of numbers of electrons vs ionization potential (the photoelectron spectrum), since $h\nu$ is a known quantity and we wish to know I_n.

A great deal of work has been carried out by UPS on gas-phase molecules. In particular, electronic energy level and vibrational fine structure information can be generated. A typical ultraviolet photoelectron spectrum is shown in Figs. 24a. Ionization of a nonbonding electron results in little change in the molecular internuclear distance and sharp peaks are obtained, whereas the removal of a bonding electron causes changes in bond length and a distribution of vibrational transitions resulting in broad peaks. Such vibrational structure has been used to study vibrational frequencies in ions, to distinguish the nature of electron bonding, and to examine the consequences of the Jahn–Teller theorem. Also found in spectra are the consequences of spin–orbit coupling, which occurs as a result of the production of a molecular ion in an orbitally degenerate state. The hydrogen halide molecules show the expected progressive increase in splitting of the orbitals of halogen lone-pair character (HF, 0.033 eV; HCl, 0.073 eV; HBr, 0.32 eV; HI, 0.66 eV). In terms of chemical applications, UPS has been employed for "fingerprint" identification in analytical chemistry, in molecular conformation studies, for examination of the orbital structure of transient species, and in correlation experiments with theoretical calculations of orbital energies.

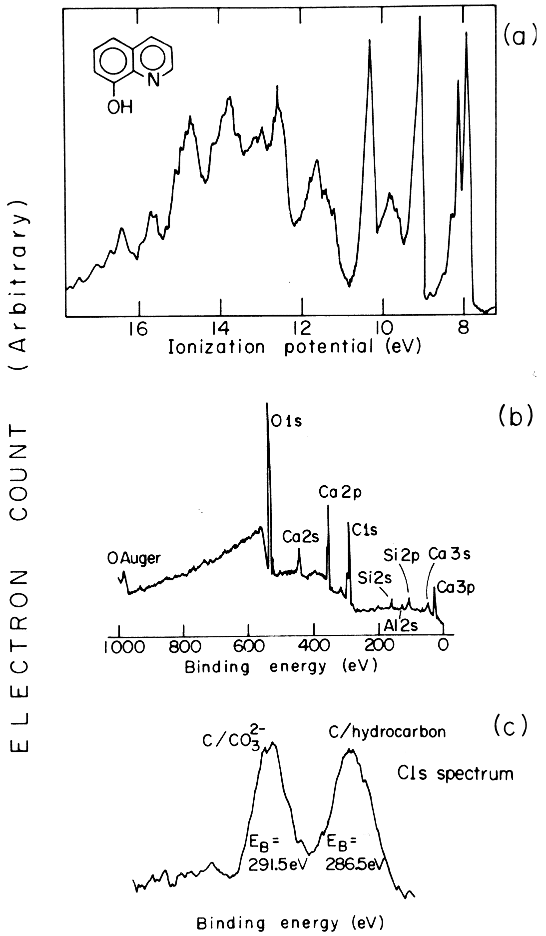


FIGURE 24 Typical vacuum ultraviolet photoelectron (a), wide-scan X-ray photoelectron (b), and narrow-scan X-ray photoelectron (c) spectra.

2. X-Ray Photoelectron Spectroscopy or Electron Spectroscopy for Chemical Analysis

The development of this technique resulted in the award of the Nobel Prize in physics to Kai Siegbahn of Sweden. Here, the sample is exposed to a beam of X-rays (although synchrotron radiation has been used), causing ejection of both valence and more deeply held (core-level) electrons. Experimentally, by far the most work has been accomplished with X-ray sources MgKα at 1253.6 eV and AlKα at 1486.6 eV, and most studies have been related to solid samples and core-level spectra. A very important feature of X-ray photoelectron spectroscopy (XPS) is that it is a surface-sensitive technique. Electrons produced by X-ray beam ionization are greatly attenuated by collision with sample atoms or molecules according to the expression

$$I(x) = I^0(x)e^{(-x/\lambda)},$$

where $I(x)$ is the electron flux emerging at the surface for a flux originating at depth x, $I^0(x)$. The parameter λ is called the escape depth and is of the order of 20 Å for electrons of kinetic energy 500 eV. Clearly, only discrete kinetic energies of electrons will be obtained from the first few atomic layers.

The binding energies of core electrons are essentially constant, within narrow limits, and are characteristic of the atom concerned. Accordingly, a sweep of the binding energy scale (computed directly from measured kinetic energy and knowledge of the source energy corrected for charging) for a particular sample leads to an elemental identification (Fig. 24b). Furthermore, the precise binding energy of a core electron varies with the oxidation state or chemical environment of an element. This is called the chemical shift effect and manifests for ionizations of the same atomic orbital from different chemical situations (Fig. 24c). Observation of this effect is particularly useful in solving structural problems in an analogous fashion to that used in nuclear magnetic resonance spectroscopy. Additional features of core-electron spectra are shake-up and multiplet structure.

There are many applications of XPS, particularly in the area of surface analysis. It has been applied in such diverse areas as polymers, catalysts, textiles, corrosion, biochemistry, environmental science, geochemistry, and archaeological chemistry. It is usually used for samples of approximately 1 cm^2 area but can sample a surface to a lateral resolution of approximately 150 μm^2 (limited by ability to focus X-rays). Multichannel analysis and long analysis times improve spectra, which often must be deconvoluted by statistical techniques to identify signals associated with closely spaced chemical shifts.

3. Auger Electron Spectroscopy

To describe Auger electron spectroscopy (AES) we must consider the fate of an atom after an electron is ejected by either X-ray radiation or electron impact. When a hole is generated in a core level, the atom must "relax," and it does this via electron movement to fill the level with vacancy. This process can result in the release of an X-ray photon (fluorescence) or can be responsible for the ejection of a secondary electron (Fig. 25a). The latter is called an Auger electron after its discoverer in the 1920s, and it is the counting and measurement of the kinetic energy of these electrons that constitute AES. As distinct from XPS, nomenclature from physics is used to describe the origin of a particular Auger electron. From elements toward the top half of the periodic table, the symbol $X_aY_bZ_c$ is used, where X is the original level where primary ionization takes place, and Y and Z represent the origins of the "down" electron and ejected Auger electron, respectively. The subscripts a, b, and c are subshell indexes, or J values. For example, we may label an electron as KL_1L_2, which means we have ionized a 1s electron, with relaxation of a 2s electron (L_1), followed by emission from L_2. AES compares with XPS in that the kinetic energies of the Auger electrons are characteristic of their elements of origin, a feature useful for analysis, and that it is a surface-oriented method for similar reasons. Also, there are other observed transitions that have to be taken into account such as shake-up and Coster–Kronig processes. These generally produce satellites to the main peaks or are responsible for the nonobservance of certain peaks, respectively.

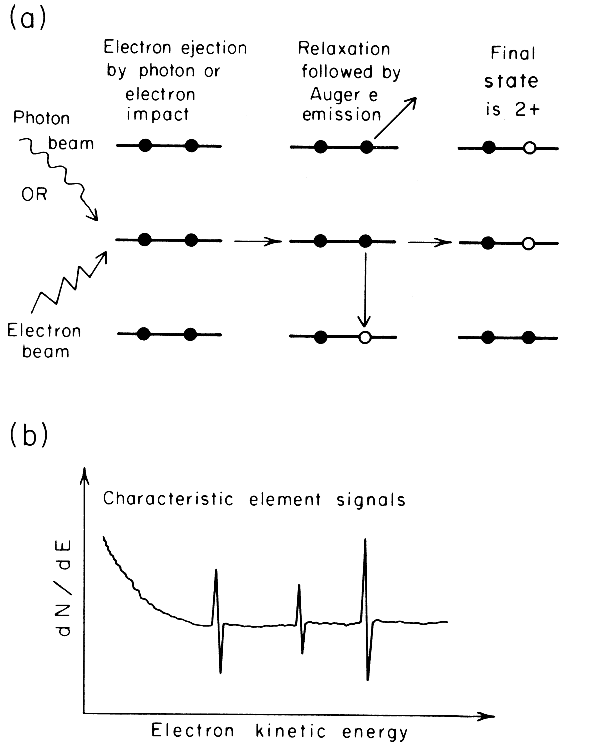


FIGURE 25 Ejection of the Auger electron (a) and typical differential Auger electron emission spectrum from a solid (b).

On a practical basis, the ionizing system is usually a monoenergetic beam of electrons. This arrangement enhances the sensitivity of the technique through many electron-impact collisions. However, this has the undesirable feature of a large diffuse background of scattered electrons, which makes it difficult to observe the discrete Auger peak in a simple manner. Accordingly, the spectrum is usually plotted in the differential mode (Fig. 25b) to pick out the identifiable Auger features. Another valuable experimental feature, associated with the narrowness of the ionizing electron beam (~1 μm), is the ability to carry out a full three-dimensional "surface" volume analysis of the

sample. Here, lateral analysis can be performed by movement of the impacting electron beam, and depth studies can be achieved by removing layers of sample by impingement of a sputtering ion beam.

Not surprisingly, AES has found tremendous use in the analysis of surfaces of samples in metallurgy and materials science. It has been used extensively in alloy analysis, metal oxidation, segregation, adsorption phenomena, catalysis, electrodeposition, corrosion, films and coatings, tribology, adhesion, and the semiconductor industry.

4. Surface Analysis

In view of the comments regarding the use of XPS and AES in surface analysis, it is appropriate to summarize this area concisely. An ideal method for surface analysis should possess the following features:

1. Is capable of monolayer examination
2. Detects elements
3. Identifies molecular species
4. Elucidates surface topography
5. Has high sensitivity
6. Has high spatial resolution
7. Is applicable to a wide range of samples
8. Does not discriminate against any component
9. Has no influence on surface composition and structure

As expected, no single technique possesses all these requirements. The battery of methods that are available is outlined in Table XII. From these data it is clear that the overall strategy is the study of information carried by emitted photons, ions, or electrons after perturbation of a

We now describe briefly the principles of a number of important methods. In secondary-ion mass spectrometry (SIMS), solids are bombarded by 1- to 30-keV ions, resulting in the ejection of substrate species as positively and negatively charged atomic and molecular particles (and neutrals). The charged species are subjected to mass spectral analysis. The method is used in both dynamic and static modes; in the latter the target is bombarded "gently," resulting in a low sputtering rate and a relatively long average lifetime of the monolayer.

TABLE XII Example Methods of Surface Analysis

Excitation or probe	Exit species and information carrier: Photons	Electrons	Ions (neutrals)
Photons	Laser optical-emission spectroscopy (LOES)	X-ray photoelectron spectroscopy (XPS)	Photodesorption (PD)
	Light (Raman) scattering spectroscopy (LS)	Ultraviolet photoelectron spectroscopy (UPS)	
	Fourier transform infrared spectroscopy (FTIR)		
	Ellipsometry (E)		
	Evanescent wave spectrofluorimetry (EWS)		
Electrons	Electron microprobe (EMP)	Auger electron spectroscopy (AES)	Electron-stimulated desorption (ESD)
	Scanning electron microscopy X-ray detection (XSEM)	Scanning electron microscopy (SEM)	
		Low-energy electron diffraction (LEED)	
		Electron-impact energy loss spectroscopy (EELS)	
Ions	Ion-induced X-ray spectroscopy (IIX)	Ion-neutralization spectroscopy (INS)	Secondary-ion mass spectrometry (SIMS)
	Proton-induced X-ray spectroscopy (PIX)	Ion-induced Auger electron spectroscopy (IAES)	Ion-scattering spectroscopy (ISS)
	Surface composition by analysis of neutral species and ion-impact radiation (SCANIIR)		Rutherford backscatter spectroscopy (RBS)
	Glow-discharge optical spectroscopy (GDOS)		
Electric field	—	Field electron microscopy (FEM)	Atom probe field-ion microscopy (APFIM)

In ion-scattering spectroscopy (ISS), low-energy ions (0.5–2 keV) bombard the surface to provide energy spectra characteristic of the masses of the scattering centers. Low-energy electron diffraction (LEED) is concerned with the impingement of electrons of energy 5–500 eV on a surface to provide a study of the surface diffraction process. In electron-impact energy loss spectroscopy (EELS), a monoenergetic beam of electrons is imposed on a surface, and the electrons obtained from the surface are examined for discrete energy losses associated with the vibrational frequencies of adsorbed molecules. Direct observation of individual molecules and atoms has been achieved by Scanning Tunnelling Microscopy (STM) and Atomic Force Microscopy (AFM), both of which bring a fine wire with a tip of atomic dimensions to the surface and observe interactions of the wire with the surface while the wire is displaced laterally by piezoelectric crystal mechanical manipulators. Finally, in Table XIII we carry out a direct comparison of the methods outlined in this article with respect to such analytical criteria as sensitivity, limit of detection, and lateral resolution. An examination of the data reveals that each method has its own strengths and weaknesses and that surface methods in general are truly complementary.

H. Chemical Sensors

Since 1975 there has been a movement toward the *in situ* analysis of specific species in complex mixtures without initial application of separation science to simplify the problem. The possible applications of such chemically selective devices, particularly those designed to quantitatively monitor organic compounds, cover a wide range of situations, including on-line monitoring and control of

TABLE XIII Performance Criteria for Methods of Surface Analysis[a]

Characteristic	AES	XPS	SIMS	ISS	LEED	EELS
Excitation beam	Electrons	X-ray photons	Ions	Ions	Electrons	Electrons
Energy (keV)	0.1–5	1–10	0.1–100	0.5–2	0.2–0.5	0.003–0.008
Diameter (μm)	25–100 →1 raster	10	10^3 →1 raster	10^3	10^3	10^3
Information depth (Å)	3–25	10–30	3–20	3–10	0–10	0–10
Monolayers	2–10	3–10	1–4	2	0–2	0–2
Detection capability						
Elements	$Z > 22$	$Z > 1$	All	$Z > 1$	Not directly	Not directly
Elemental sensitivity range	10	10	10^4	10	—	—
Isotopes	No	No	Yes	Restricated	No	Restricted
Chemical valence	Special cases	Yes	Indirect	No	No	Yes
Organics	No	Yes	No	No	No	No
Beam damage	Small	Occasionally	Yes, dynamic; uppermost layer, static	Small	Small	No
Lateral resolution	<1μm possible	~1.0 mm	<1 μm possible	10 μm	2.0 mm	1.0 mm
Detection limits						
"Surface" (g cm^{-2})	10^{-10}	10^{-9}	10^{-13}	10^{-10}	10^{-10}	10^{-10}
Bulk (atomic fraction)	10^{-3}	10^{-3}–10^{-2}	10^{-3}–10^{-4}	10^{-2}	—	—
Advantages	Sensitive to low-Z elements; minimal matrix effects; high lateral resolution	Information on chemical bonds; no beam damage	Detection of all elements and isotopes; good detection sensitivity; high lateral resolution	Outermost atomic layer analysis	Atomic structure of "ordered" surface	Direct information on interaction of adsorbate
Disadvantages	Difficult to quatify; no H, He detection	No lateral resolution; slow profiling; no H detection	Difficult to quantify matrix effects	Low sensitivity, poor lateral resolution; slow profiling	No elemental analysis pattern; often difficult to interpret; long analysis time	No elemental analysis; long analysis time

[a] For definitions of acronyms, see Table XII. [From Thompson, M., Baker, M. D., Christie, A., and Tyson, J. F. (1985)."Auger Electron Spectroscopy," Wiley-Interscience, New York.]

industrial chemical processes, long-term remote environmental monitoring, automation of clinical analyses, *in vivo* drug monitoring, and feedback control of artificial organs.

All chemical sensors employ a basic union between a selective receiving site for ionic or molecular binding and a transducer that is capable of translating a chemical binding into a useful analytical signal. Transducers can be divided into four categories: semiconductors, conventional ion-selective electrodes, optical devices, and piezoelectric devices. Other than the ion-selective electrodes described previously, the semiconductor-based gas sensors are the best known chemical sensors. These devices usually operate at elevated temperatures, allowing certain gases to interact chemically with the semiconductor surface to alter its electrical characteristics. Such devices may be selective to certain gases such as CO_2 and O_2 or to certain classes of molecules such as hydrocarbons and may take the form of chemiresistors or thin-film metal oxides.

Better selectivity can be attained by the use of specialized chemical receptors. The ion-exchange chemistry of ion-selective membranes is one manner in which selectivity is enhanced for simple inorganic ions and gases. Nature provides the materials for the complicated task of selectively complexing biochemicals through molecular recognition processes. The selectivity of these receptors is based on the tertiary structure of proteins and polypeptides and on nucleic acid complementary pairing. Enzyme–substrate, antibody–antigen, lectin–saccharide, and hormone receptor–hormone selective reactions have all been employed in conjunction with "building block" construction and a variety of transducers to produce sensitive and selective devices. Such selective chemistry has also found wide application in techniques such as immunoassay. The development of catalytic antibodies (synzymes) and genetic engineering to express proteins of DNA of interest for chemical selectivity and the introduction of artificial receptor sites and selective surfaces have substantially broadened the commercial potential for long-term development of a wide range of chemical sensors.

1. Electrochemical Devices

These devices are based on the measurement of either electrochemical potential or faradaic current associated with redox reactions at an electrode. They are particularly suitable for enzyme–substrate receptor systems by virtue of the ionic products often produced in such reactions. The sensing membranes of the ion-selective electrodes previously described have been combined with semiconductor devices for miniaturization, low-impedance output, signal amplification, and capability of on-chip processing. The ion-sensitive field effect transistor (ISFET) is based on replacement of the conventional transistor gate with the ion-selective membrane. The variation of potential of one surface of this membrane can control any electronic current through the transistor known as the drain current. A similar arrangement employing an enzyme–substrate receptor system at the gate also provides for ionic charge control of transistor electronic current conduction, resulting in a chemically sensitive field effect transistor (CHEMFET).

Multiple enzyme systems, where one enzyme produces an electrochemically inactive product that is consumed as a substrate by another enzyme to form an active product, have been successfully used to extend enzyme selectivity. The selectivity of immunochemical systems has been employed by implementation of enzyme-linked assays. Direct coupling of redox relay centers of enzymes to conductive electrodes has been achieved by a technique known as molecular wiring and avoids the indirect analysis of products of enzyme–substrate reactions. This fast and sensitive technique measures current flow and is commercially available.

2. Optical Devices

In analogy to electrodes, chemically selective optical devices have been termed optrodes. These systems employ absorption and luminescence strategies and make use of waveguide fiber-optic and laser technology for miniaturization. Light in the ultraviolet or visible region of the electromagnetic spectrum can be passed through an optical fiber to a remote reaction vessel. Detection may depend on observation of light passing through the reaction zone, or being transmitted toward a detector by another fiber, or by the evanescent wave phenomenon encountered in techniques such as attenuated total reflection or total internal reflection fluorescence spectroscopy. In contrast to electrochemical devices, these systems offer advantages such as insensitivity to electrical interference, elimination of reference electrodes, possibility for change of reagent phase, distributed multisensor operation with the use of one optical detector (spectrophotometer), multiwavelength analysis, time-resolved analysis, and a complementary range of analytes that may not be electrochemically active. One example is a pH sensor, based on optical absorption changes of an indicator dye trapped in a gel matrix in a cell at the tip of a fiber-optic system. Ratio methods of quantitative analysis and multiwavelength analysis for selective observation of two or more species are preferred for elimination of background drift and interference problems. A form of evanescent wave spectroscopy in which the electromagnetic field is coupled into the conduction band of a thin metal film is currently being commercialized. This technique, known as Surface Plasmon Resonance Spectroscopy (SPR), is very sensitive to any organic reagents located on the exterior of the metal film

and can be used to picomolar detection limits for many immunochemical analyses.

3. Piezoelectric Devices

It is well known that certain crystalline substances such as quartz generate an electrical voltage when subjected to physical compression. Similarly, it is possible to induce a constant frequency of physical oscillation of the crystal planes when a regular alternating voltage is applied to such a crystal. Piezoelectric crystals coated with thin films of selective adsorbent have found application as conventional gas detectors when employed as physical oscillators. The mass changes caused by the adsorption of gases to the crystal cause the oscillation frequency to change, Δf, in proportion to adsorbed mass as

$$\Delta f = -2 f_q^2 \big/ (P_q V_q) m_f,$$

where f_q is the quartz crystal frequency, P_q is the quartz density, V_q is the velocity of the wave in the quartz, and m_f is the mass per unit area for the deposited matter on the quartz surface. It has been demonstrated that mass responses and also microviscosity responses associated with surface reactions can be obtained from enzyme–substrate and immunochemical complexation on bulk acoustic wave piezoelectric quartz crystals. Furthermore, both bulk acoustic and surface acoustic wave devices can be used directly in aqueous media in certain configurations for quantitative analysis.

III. COMPUTERS IN ANALYTICAL CHEMISTRY

A. Instrument Control and Data Handling

Since 1975 there has been a tremendous revolution in capability and cost of computers and microprocessors, resulting in the incorporation of such technology into almost every modern analytical instrument available today. Computers are fundamental to two areas of analytical chemistry: (1) the applications of mathematics to analysis (chemometrics), encompassing such topics as factor analysis, multiple regression, pattern recognition, optimization, and statistics, and (2) interactions with analytical instruments. The latter can actually be considered to be either passive or active interaction, as determined by computer control of instrument operation.

1. Data Acquisition and Processing

This area usually implies a passive function involving the actual collection of data, simple calculations, data averaging, statistical analysis, integration, and data storage. Often these functions can be handled by a micro- or minicomputer with a minimum of 8-bit word lengths, though greater word length and greater capability of microprocessor addressing allow for increased processing speeds and greater on-line data storage capability. Data acquisition is usually performed by specialized analog-to-digital circuit packages, which may receive polling requests and timing signals from the microprocessor. Two common devices for transducing analog instrument measurements into digital form are known as counter converters and successive approximation counters. Some instruments acquire complicated and extensive data sets at rapid rates. Such data may be difficult to interpret manually, but with the aid of a reference library, aspects such as compound identification can readily be established by computer. Such reference libraries are used extensively for infrared and mass spectral analysis.

2. Microprocessor Control

Due to its great speed, a computer can often control an instrument more efficiently than is possible by manual operation. Such active instrument control must occur in at least two steps. The microprocessor must first ascertain the state of some variable, and then action must be taken to activate an instrument control to adjust the value of the variable. Many instruments employing microprocessor control contain more than one processor and can perform a series of complicated feedback functions. These processors are considered to be "dedicated" to a particular series of tasks and receive their operating commands from a hard-wired instruction set as well as the instrument operator. The simplest microprocessors operate with instruction sets based on 8-bit word lengths. Memory increments of 256 words are available for such systems, and often only one or two of these memory increments are required for instrument control. Control usually takes the form of activating a standard electrical relay device (which may drive a motor or switch) by toggling an active line connected to the relay and controlled by the microprocessor between electrical ground and a standard 5-V high level. More complicated control systems involve more control relays and necessitate the availability of more control lines and microprocessors. The time-sharing capabilities of a minicomputer are often used in controlling a number of microcomputers or microprocessors in complex instruments. Significant advances in speed and therefore "real time" complexity of instrument operation have been achieved as new 16- and 32-bit microprocessors have entered the market. The advantages of speed and increased memory have aided in the development of expert systems and parallel processing. A revolution in instrument control and data processing is

currently developing with the advent of parallel processing and neural networks for interactive "learning."

3. Fourier Transform Analysis

One application of high-speed computers to data analysis is often found in spectrophotometric applications, such as infrared and nuclear magnetic resonance techniques. Samples can be irradiated with broad ranges of frequencies from the appropriate regions of the electromagnetic spectrum and will absorb certain discrete frequencies dependent on sample chemistry. Each independent frequency that can be observed (resolved) in the range of energies employed can be represented as a sinusoidal oscillation. The simultaneous superpositioning of all the available frequencies produces both constructive and destructive interference, resulting in a well-defined complex waveform pattern. Interaction of the sample with discrete frequencies will alter the waveform pattern, which will then contain the analytical interaction information in the form of a time "domain." This can be converted to a conventional frequency-domain spectrum by the fast Fourier transform algorithm, so that individual frequencies that make up the superimposed waveform can be individually identified and plotted in conventional formats. Data must be sampled and digitized at a rate at least twice the value of the ratio of the range of frequencies encountered divided by the frequency resolution desired. The major advantage of this technique is that all frequencies are simultaneously measured, and a complete conventional spectrum can be constructed in seconds for any one measurement. Since these spectra are digitized and contain frequency reference information, it is possible to sum sequential spectra to improve signal-to-noise ratio. Signals increase linearly with spectral addition, while noise increases as the square root of the number of spectra that are combined.

B. Chemometrics

The term *chemometrics* describes the interface between analytical chemistry and applied mathematics, where mathematical and statistical methods are employed to maximize information quality in a chemical experiment. Most chemometric methods involve matrix algebra, which is efficiently handled by computer, and numerous programs are presently available. A number of reviews have been written on this broad subject area, which includes such topics as statistics, modeling and parameter estimation, resolution, calibration, signal processing, image analysis, factor analysis, pattern recognition, optimization strategies, and artificial intelligence. Appropriate topics can be chosen to optimize an analysis at each level of experimentation, including sampling, measurement, data processing, resultant conversion, and information organization. Software packages for these purposes are commercially available.

1. Common Chemometric Methods

The three most commonly used chemometric methods are discussed in the following subsections.

Multiple regression analysis. This is suitable for data modeling and expresses data as a simple equation. The process begins with experimentation to produce a vector of measured data known as the "dependent" variables. Then a limited number of "factors" are considered to be significant for the determination of data values, and these "independent" variables are used to prepare a model for the data. Finally, coefficients, as shown below, are calculated by least-squares analysis to represent the significance or weighting of the independent variables. The result is a calculation of "regression coefficients" to prepare a mathematical model that is suitable for preditions,

$$d = c_1 i_1 + c_2 i_2 + \cdots + c_n i_n,$$

where d represents the dependent variable, c represents the regression coefficient, and i represents the independent variable.

Factor analysis. This method is used to interpret underlying factors responsible for data and is one of the most versatile chemometric methods. Factor analysis provides a purely mathematical model prepared from abstract values, which are related to a data matrix as follows,

$$D = RC,$$

where D represents the data matrix and R and C represent factors for each row and column. The factors are mathematically transformed so that their significance can be interpreted with respect to the data. This results in the establishment of the number of significant factors and assists in the correlation of data and the application of physical significance to the factors.

Pattern recognition. This procedure allows the classification of a species to be made on the basis of a series of measurements that establish a pattern. Procedurally, a matrix describing the patterns of a number of species is constructed. Then a decision vector is designed by the use of standards to divide the patterns into discrete classifications, resulting in a mathematical form,

$$p = V_1 d_1 + V_2 d_2 + \cdots + V_n d_n,$$

where p represents a set of patterns, V represents components of the decision vector, and d represents the data

for the species of interest. The pattern p is assigned and values of d are determined so that a decision vector **V** can be calculated. When the calculated vector **V** is mathematically combined with a new set of experimental data, a pattern p is calculated for the experimental species and can be fitted to previous classifications.

2. Optimization by the Simplex Method

Numerous mathematical techniques exist for solving a series of simultaneous equations given defined boundary constraints in order to maximize or minimize a particular parameter. The general acceptance and implementation of techniques such as linear programming attest to the power of optimization strategies. The simplex method is an "evolutionary operations" method that has been used systematically in many problems. A simplex is a geometric figure whose vertices are defined by the number of experimental parameters plus 1. Each point of the simplex represents the actual measured analytical response at a set of chosen experimental parameters. Represented in some n-dimensional space, one vertex of the simplex always represents the case of worst response in the experimental. A mirror reflection through a symmetry plane away from the point of worst response (assuming the response will be greater at a point opposite to the worst case) generates another simplex. An experiment is then performed using the new parameters to determine which vertex represents the new worst case response. A reiteration process following four well-defined rules allows movement along the "response surface," resulting in eventual convergence to the optimal experimental conditions.

The basic rules are as follows:

1. Rejection of the point with the worst result is followed by replacement with its mirror image across a line or plane generated by the other remaining points.
2. If the new point has the worst response, the previous simplex is regenerated and the process applied in rule 1 is repeated for the second worst case point.
3. If one point is common to three successive simplexes, it represents the optimum, provided that the point represented the best response in each case. If this is false, the entire process must be repeated using new initial starting points.
4. Boundary conditions are defined so that if a point falls outside accepted bounds, it is assigned an artificially low value, which forces the simplex to move into the useful calculation area.

A variation of the latter optimization procedure known as "modified simplex optimization" has evolved to eliminate the limitations imposed by the simplex method. The method originally described must employ steps of fixed size, which can result in excessive experimentation when step size is small or in poor precision for large steps. A more efficient solution employs variable step size throughout the entire procedure, allowing expansion (acceleration) of the simplex in favorable directions and contraction in zones that produce poor results. The distance to be moved is controlled by constant arbitrarily chosen multiplication factors, which are multiplied with the distance of movement obtained on reflection. Eventually, the simplex contracts as movement to the optimum occurs. The process halts when the distance of movement has dropped below some predetermined value, which is generally governed by experimental uncertainty or time limitations. Certain difficulties exist in the application of simplex methods when considering error sources:

1. The method cannot be used if discontinuous variables are chosen.
2. Movement to a local optimum may occur if numerous optima exist.
3. Parameters must be judiciously selected to ensure that nontrivial analyses occur.
4. As many significant parameters as possible should be included in the simplex so that no important factors are overlooked. This subsequently increases the experimental work for each step in the simplex generation.

3. Selectivity vs Specificity

A definition of terminology has been attempted, where the upper limit of the concept of selectivity implies specificity. A *fully selective* system can measure one component in the presence of many others, while *fully specific* implies that in all situations only one component is measured and other components in the experiment do not produce any signal. A nonselective system produces an analytical signal due to all components in the experiment. For any of these cases, the measured signal x is a function of concentration c of the available component and is related to the latter by a normalization parameter γ, where

$$x = \gamma c.$$

The element γ is determined by the sensitivity parameter dx/dc, and the sensitivity of the method is numerically determined by the value of γ. This example can readily be expanded to consider a multicomponent case, where x, γ, and c become matrix representations and a partial sensitivity $\partial x/\partial c$ is employed. A mathematical rearrangement of the γ_{ij} matrix to place the largest γ value in each row on the main diagonal results in a useful "calibration

TABLE XIV Areas of Future Instrumental Development

Area	Anticipated development
Electroanalytical methods	Molecularly designed electrodes, speciation
	Microelectrodes, biological probes
	Electrochemical detectors, combination techniques
Chromatographic and separation methods	Specialization methods for difficult separations (maximization of column efficiency and methodology presently attained developments in supercritical fluid chromatography, separation of isomers and chiral species)
Spectrochemical methods	Higher yield ion sources for mass spectrometry
	Hyphenated techniques, gas or liquid chromatography with inductively coupled plasma emission spectroscopy
	Minimization of sample preparation
Chemometrics	Computer-assisted data manipulation
	Expert systems and "intelligent" instruments
	Resolution improvement programs for chromatography to speed analyses
	Pattern recognition systems
Surface science	Angle-resolved electron spectroscopy
	Molecular information, conformations
	Discrete molecular and atomic resolution, e.g., for direct sequencing of DNA

matrix." If the corresponding "analytical matrix" of x_{jk} is considered, the following observations apply:

1. For a determinant of $\det(\gamma_{ik}) = 0$, no mapping from the calibration to the analytical matrix exists.
2. Selectivity can be given a value θ, where

$$\theta = \min_{i=1 \to n} \frac{\gamma_{ii}}{\sum_{n=1}^{k} \gamma_{ik} - \gamma_{ii}} - 1,$$

   A system is fully selective when $\lim \theta \to \infty$, but no selectivity exists for θ values that are small compared with zero.
3. Specificity is a special case of selectivity and generates a single nonzero element (in the calibration matrix), which lies on the diagonal.
4. A partial specificity can be defined when only the diagonal element in one row can practically be considered to have a nonzero value. A grade of specificity is numericaily determined by

$$\theta_a = \frac{\gamma_{aa}}{\sum_{k=1}^{n} \gamma_{kk} - \gamma_{aa}} - 1.$$

IV. FUTURE PERSPECTIVES

Naturally, it is difficult to predict the evolution of a discipline as diverse as analytical chemistry. Table XIV indicates a summary of short-term future directions as garnered from the present interests and activities of researchers in the field. One dramatic, rapid change is the movement away from sampling technology. Direct *in situ* measurement technology is being emphasized in many areas of analytical chemistry and eventually may largely supercede the need to define statistically valid samples from a bulk sample material.

Probably the analytical instrumentation of the future will become more and more automated, but until artificial intelligence makes its debut in instrumentation, the chemical knowledge of the analyst will always be of paramount importance. The intelligent application of any analytical technique will continue to require a good understanding of basic physical and chemical theory and a knowledge of practical experimental limitations.

SEE ALSO THE FOLLOWING ARTICLES

Atomic Spectrometry • Chromatography • Distillation • Electrochemistry • Elemental Analysis, Organic Compounds • Gas Chromatography • Infrared Spectroscopy • Liquid Chromatography • Mass Spectrometry • Nuclear Magnetic Resonance • Organic Chemistry, Compound Detection • Raman Spectroscopy

BIBLIOGRAPHY

Bard, A. J., and Faulkner, L. R. (2001). "Electrochemical Methods," Wiley, New York.

Christian, G. D. (1994). "Analytical Chemistry," 5th ed., Wiley, New York.

Dean, J. A. (1995). "Analytical Chemistry Handbook," McGraw-Hill, New York.

Elving, P. J., and Kolthoff, I. M. (eds.) (1978). "Treatise on Analytical Chemistry," 2nd ed., Wiley-Interscience, New York.

Kellneer, R., Mermet, J.-M., Otto, M., and Widmer, H. M. (1998). "Analytical Chemistry," Wiley-VCH, Weinheim.

Lindon, J. C., Tranter, G. E., and Holmes, J. L., eds., (2000). "Encyclopedia of Spectroscopy and Spectrometry," Academic Press, San Diego.

Meyers, R. A., ed. (1998). "Encyclopedia of Environmental Analysis and Remediation," Wiley, New York.

Meyers, R. A., ed. (2000). "Encyclopedia of Analytical Chemistry Applications, Theory and Instrumentation," Wiley, Chichester.

Skoog, D. A., West, D. M., and Holler, J. F. (1997). "Fundamentals of Analytical Chemistry," 7th ed., Saunders, Philadelphia.

Skoog, D. A., Holler, J. F., and Nieman, T. A. (1998). "Principles of Instrumental Analysis," 5th ed., Harcourt, Philadelphia.

Sorum, C. H., and Lagowski, J. J. (1991). “Introduction to Semimicro Qualitative Analysis,” 7th ed., Prentice-Hall, Englewood Cliffs, NJ.

Strobel, H. A., and Heineman, W. R. (1989). “Chemical Instrumentation: A Systematic Approach,” 3rd ed., Wiley, New York.

Svehla, G., and Suehla G. (1996). “Vogel’s Qualitative Inorganic Analysis,” 7th ed., Addison-Wesley, Reading, MA.

Wilson, I. D., Adlard, T. R., Cooke, M., and Poole, C. F., eds. (2000). “Encyclopedia of Separation Science,” Academic Press, San Diego.

Atomic Spectrometry

Vahid Majidi
Chemistry Division, Los Alamos National Laboratory

I. Atomic Spectrometric Techniques
II. Optical Detection
III. Atomic Emission Spectrometry
IV. Atomic Absorption Spectrometry
V. Atomic Fluorescence Spectrometry
VI. Detection of Elemental Ions
VII. Comparisons of the Atomic Spectrometric Techniques

GLOSSARY

Analyte Element whose concentration is to be determined in a matrix.

Atom cell Device that provides a sufficiently high temperature to break down the sample into its constituent atoms.

Differential pumping Evacuating two chambers, adjoined by a conductance limited orifice, with the use of either one or two independent vacuum pumps. This allows for maintaining two connected chambers at different pressures with respect to one another.

Interference Increase or decrease in the size of the signal obtained from the analyte as a result of the presence of some other known or unknown component in the matrix.

Matrix Components of a sample other than the element whose concentration is to be determined.

Multi-element Multi-element analysis is the simultaneous or rapid sequential determination of the concentration of several elements in a matrix.

Standard Sample that contains the analyte at a known concentration from which a signal can be obtained that can be compared with the signal from a sample of unknown analyte concentration. From this, the concentration of the analyte in the unknown sample can be determined.

ATOMIC SPECTROMETRY is a series of analytical techniques that are used to determine the identity and concentration of elements in any sample. While historically the detection of analytical signals was exclusively done through measurement of photons, during the past two decades atomic spectroscopists have successfully integrated the measurement of ion current into the suite of detection technologies known as atomic spectrometry. Photon-detection-based techniques include atomic emission, atomic absorption, and atomic fluorescence spectrometry. Ion-detection technologies include laser-enhanced ionization and elemental mass spectrometry (MS). Absolute analyte masses at the nanogram

level (10^{-9} g) can be determined in samples of a few micrograms (10^{-6} g) by electrothermal vaporization (ETV) atomic absorption. Solution concentrations at the nanogram (10^{-9} g) per milliliter level can be routinely measured by most of these techniques. Conversely, using plasma mass spectrometry, absolute mass detection limits of a few femtograms (10^{-15} g by ETV–MS) or measuring solution concentrations of parts per quadrillion (10^{-15} g/ml) are not uncommon. All the techniques use the phenomenon of electronic excitation. This involves transitions among the outer electrons in the atomic orbitals, with accompanying ionization, emission, or absorption of radiation. In photon-based detection, the intensity of these emissions and absorptions can be related directly to the concentration of an element in a sample. For ion-based detection, the obtained electric current is directly proportional to the analyte concentration.

I. ATOMIC SPECTROMETRIC TECHNIQUES

There are three traditional analytical atomic spectrometric techniques: atomic emission spectrometry (AES), atomic absorption spectrometry (AAS), and atomic fluorescence spectrometry (AFS). In all three cases, the analytical signal is generated by photons. More recently, however, the detection of resulting ions from an electronic excitation has led to superbly sensitive detection of elements. It is possible to distinguish among these techniques by considering the excitation mechanism and the way the analytical signal is detected. The general schematic representation for various types of atomic spectrometry is shown in Fig. 1, where two-level electronic systems are depicted. The ground state is labeled 0, and the excited state is labeled 1. In general, the transition originating from or terminating to the ground state (or another of the lowest lying excited states) is most commonly used for atomic spectrometric measurements (Ingle and Crouch, 1981).

A. Atomic Emission Spectrometry (AES)

The most basic measurements in analytical atomic spectrometry can be traced back to Thomas Melville, who in 1752 reported his observations on spectra of mixtures of alcohol with sea salts (Laitinen and Ewing, 1981). The simplest example of atomic emission is the experiment of putting table salt (sodium chloride) into a flame, which generates a yellow color. Although the example is a simple illustration, the actual events that ultimately lead to the yellow plume are due to a complex series of chemical and physical processes outlined in Fig. 2. Sodium chloride is initially volatilized and ultimately decomposed into its constituent atoms (sodium and chlorine) by the heat of the flame. The thermal energy from the flame excites the gas-phase sodium atoms from the ground state (0) to an excited state (1). The efficiency of this excitation depends on the gas-phase temperature (the wavy line in Fig. 1A represents the thermal excitation processes). The sodium atom in the excited state then spontaneously decays back down to the ground state, either by collisional transfer of energy to other species or by atomic emission (loss of energy by emission of a photon). This is illustrated by the straight line from the excited state down to the ground state in Fig. 1A. The wavelength of this emitted photon depends on the energy difference between the excited state and the ground state.

The amount of light that is emitted from the analyte is proportional to the number of atoms in the flame or plasma. Hence, atomic emission spectrometry can be both a qualitative analytical technique, in which the identity of a metal is revealed by the observed color (wavelength), and a quantitative analytical technique, in which the intensity of the light emitted from the analyte is a function of the number of atoms.

The experimental arrangement involved in an AES measurement is shown in Fig. 1E. The hot analyte environment, which is able to break down and excite atoms, is called an *atom cell*. The atom cell can be a flame, plasma, a heated graphite tube, or any other environment where the analyte is observed in a spatially confined arrangement. In Fig. 1, the detector box is used to represent a detection system, which is able to identify the wavelength and measure the intensity of the emitted radiation. The experimental arrangement is the simplest of the three optical atomic spectrometric techniques.

B. Atomic Absorption Spectrometry (AAS)

Figure 1B illustrates the principle of atomic absorption spectrometry. The excitation from the ground state to the upper state is by the absorption of light energy. An atom cell (a flame or a heated graphite tube) is used to decompose compounds, but the energy for excitation is drawn primarily from a light source such as a white light source (which emits all wavelengths) or a hollow cathode lamp (which emits narrow spectral lines). In AAS, the detection system looks at the light source directly (Fig. 1F) and sees an intensity of the light source (I_0) before any atoms are present within the atom cell. When atoms are introduced into the atom cell, they absorb some of the radiational energy (solid arrow in Fig. 1B) and are excited from the ground state to the upper state. The detection system sees this absorption as a reduction in the intensity of the light source from I_0 to I (Fig. 1F). The ratio I/I_0 is called the

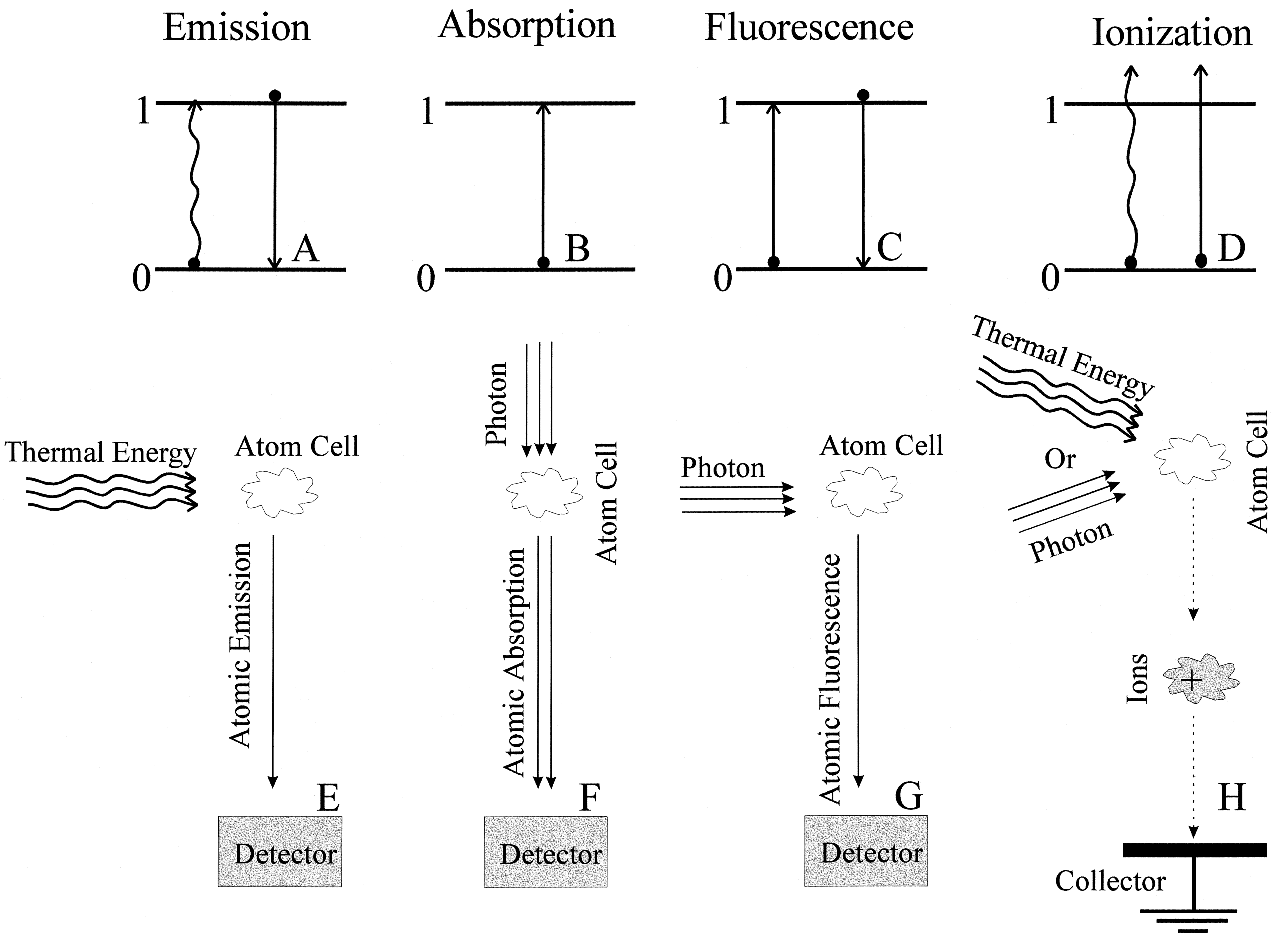


FIGURE 1 A simple two-level energy diagram representation for excitation and relaxation mechanisms in the atomic emission process (A), atomic absorption process (B), atomic fluorescence process (C), and ionization process (D). The resulting analytical signals from the above processes are measured by experimental arrangements depicted in E, F, G, and H, respectively.

transmittance and is related to the concentration of the absorbing analyte atoms. In short, the greater the amount of light that is absorbed, the more atoms are present in the atom cell.

C. Atomic Fluorescence Spectrometry (AFS)

In AFS, the excitation from the ground state to the upper state (Fig. 1C) is accomplished by absorption of a photon, and the measurement is made by detection of radiation emitted by the atom as it relaxes back from the excited state to the ground state. This re-emitted radiation is no different in character from the radiation emitted in the atomic emission technique. In AES, the atoms are excited only by the collisional energy inherent in the high temperature of the atom cell, while in AAS and AFS the atoms are excited primarily by radiation from an external light source.

The experimental arrangement for AFS is illustrated in Fig. 1G. A light source is used to irradiate the atom cell to yield excited atoms. The direction of observation of the fluorescence is at right angles to the angle of incidence of the excitation source on the atom cell to avoid looking directly at the light source. Once again, a higher population of atoms in the atom cell produces a greater fluorescence signal.

D. Detection of Elemental Ions

When the deposited energy into an atom cell is sufficiently high, elemental ions are produced. The atoms absorb enough thermal or radiational energy to promote excitation of an electron above the bound state.

In laser-enhanced ionization, laser sources are used to selectively produce the ion from a specific element. This may be done either through resonance transitions or by exciting an electron to a bound state which is then ionized through subsequent collisions. This mode of ionization will yield a unique ion from an ensemble of different

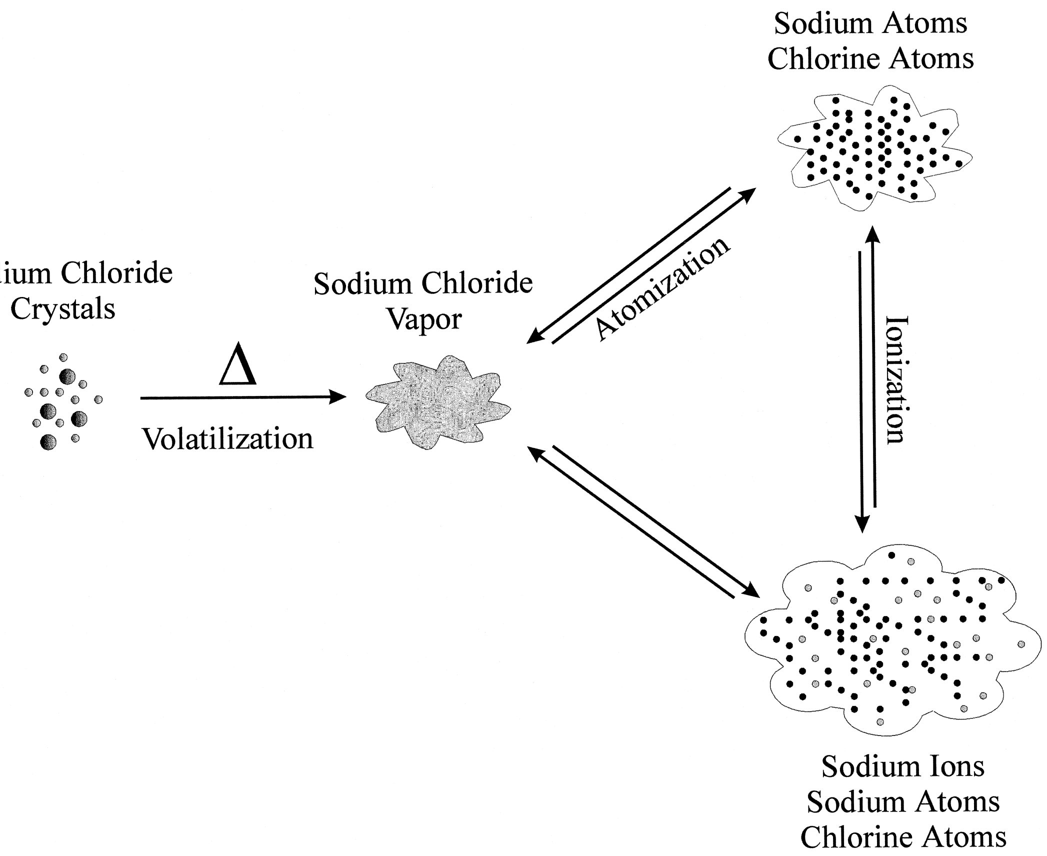


FIGURE 2 Sequence of events that occur in flame during the conversion of crystalline salts into gas-phase atomic and ionic species for spectrochemical analysis.

elements and as such an undiscriminating detector (e.g., a charge collector) may be used to selectively identify a particular element (the selection is done through the unique excitation wavelength). Conversely, in elemental mass spectrometry, thermal sources (e.g., inductively coupled plasma) are commonly used to ionize most elements present within the plasma, to varying degrees. Therefore, a mass analyzer is needed prior to the detection to provide quantitative information for a given element without interference from all other ions present. Regardless of the mode of ionization or detection, the ion current generated from an atom cell is proportional to the total amount of analyte initially present. The process of ionization and detection is shown schematically in Figs. 1D and H.

II. OPTICAL DETECTION

In atomic absorption, emission, and fluorescence spectrometry, the detection and characterization of photons are the fundamental measurements that ultimately lead to the determination of elemental identity or the quantitative evaluation of a given analyte. Therefore, for a complete spectrochemical analysis, one must be able to measure the photon intensity (to extract quantitative information) and characteristic wavelength (to extract qualitative information). The optical information encoded in any of the above spectrometric processes must be decoded in terms of intensity and wavelength.

A. Wavelength Selection

The light emanating from an atom cell or the light used in atomic absorption experiments is generally composed of several different wavelengths. The simplest approach for wavelength discrimination is to use an optical filter. But, to distinctly identify elements by their characteristic spectra and to quantify the amount of analyte present by its signal intensity, a wavelength-selection device is used. These devices are known as monochromators or polychromators, depending on their ability to display a single wavelength or multiple wavelengths at one time. Schematic representations of several different types of monochromators and polychromators are shown in Fig. 3.

Typically, the light from an experiment is focused on the entrance slit of the monochromator. The light on the

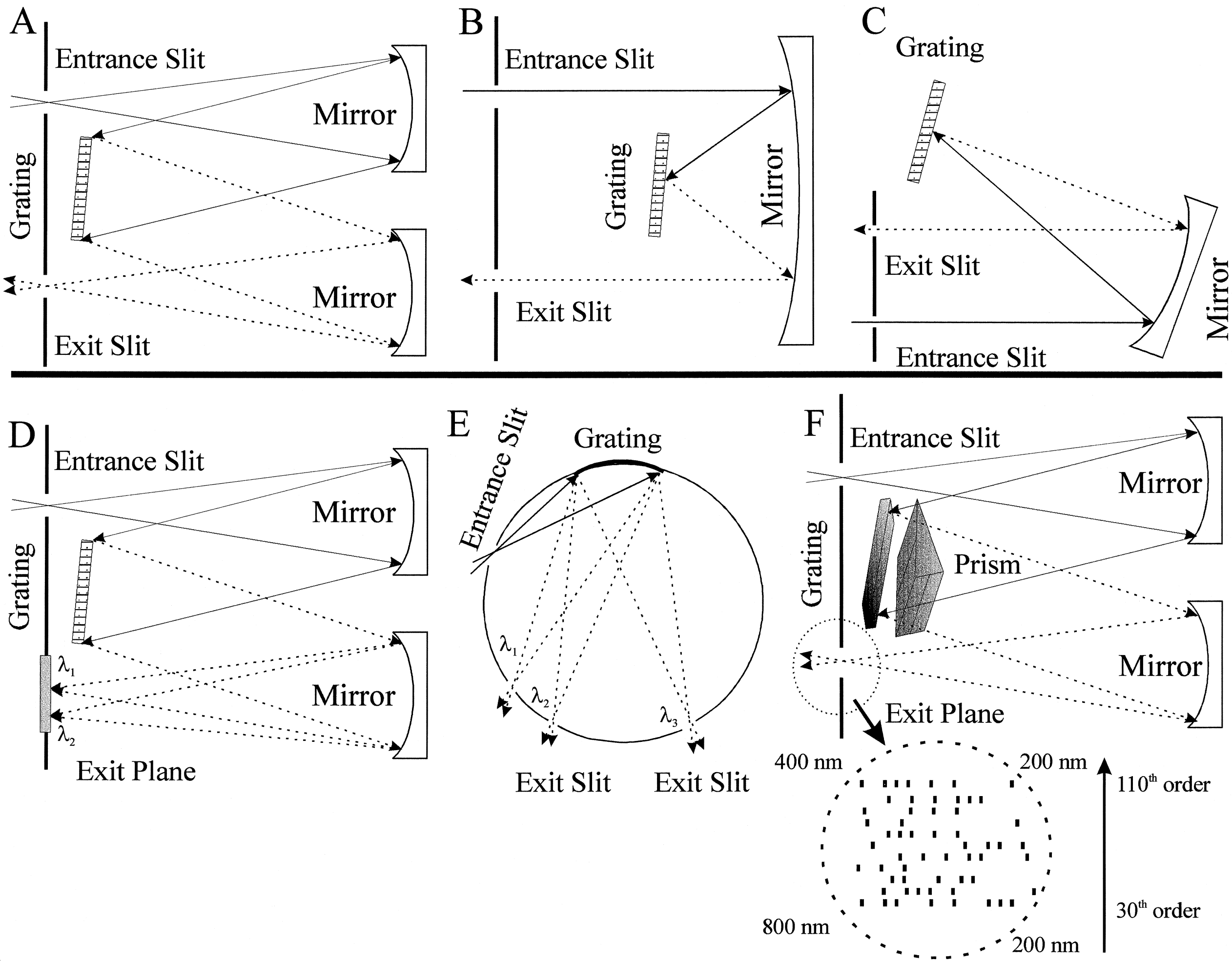


FIGURE 3 Schematic representation of the most frequently used grating-based monochromators (A, B, C) and polychromators (D, E, F). A and D are known as Czerny–Turner design. Fastie-Ebert (B) and Littrow (C) have similar designs in that both use a single reflective mirror. A Rowland circle polychromator is depicted in E, and F represents an Echelle spectrograph.

slit is then collected by a collimating mirror and converted to parallel light rays that are directed toward a wavelength-dispersion element (diffraction grating or prism). The original light is now separated into discrete wavelengths, which are imaged by the mirror on the exit plane of the monochromator. A slit is used to allow for propagation of a unique wavelength out of the monochromator and into the detector. In most instances, a monochromator can be converted to a polychromator by simply removing the exit slit and replacing the traditional detector by an area detector (see next section). On the other hand, any polychromator can be used as a monochromator if only a single output channel is used during an experiment. Although many wavelength-selection devices are available commercially, the designs shown in Fig. 3 are among the most popular. The Czerny–Turner monochromator is extremely popular because of its compact design, while the Echelle system is the polychromator of choice due to its superb spectral resolution.

B. Photon Detection

Historically, photographic films and plates have been used at the exit plane of the monochromators to detect the spectral patterns from analytes. The main advantage of this approach is its inherent multi-wavelength detection ability. Because photographic films and plates are area detectors with fantastic spatial resolution, the entire spectrum from a sample may be recorded on a 30-cm plate, which ultimately results in simultaneous multi-elemental analysis. Unfortunately, the dynamic range of photographic emulsion is limited and the response function is typically nonlinear. Furthermore, the overall process is extremely time consuming and chemically intensive (need for a darkroom,

developing time, and chemicals for the development process). Today, photomultipliers (PMT), photodiodes, and array detectors completely dominate the photon detection arena. Schematic representations for a photomultiplier and a charge injection device (a two-dimensional array detector) are shown in Fig. 4. Photomultipliers and photodiodes are used as single-channel detectors with both monochromators and polychromators. Array detectors are used for simultaneous detection over a given spectral range.

In photomultipliers, a photon strikes the photoactive material on a photocathode, resulting in ejection of an electron from the surface. The initial electron is accelerated toward the first dynode, which upon collision releases several secondary electrons. This process is repeated by using several additional dynodes to yield an overall signal amplification of 10^6. Solid-state detectors (photodiodes and charge injection devices), however, do not intrinsically amplify the signal. In these devices, the initial photon results in a single charge-hole separation or single charge storage. In all cases, the signal from the detector is electronically conditioned (e.g., amplified, filtered, time-gated, etc.) and is converted into an analog or digital format to be used with a display device (e.g., computer, analog gauge, digital display, etc.).

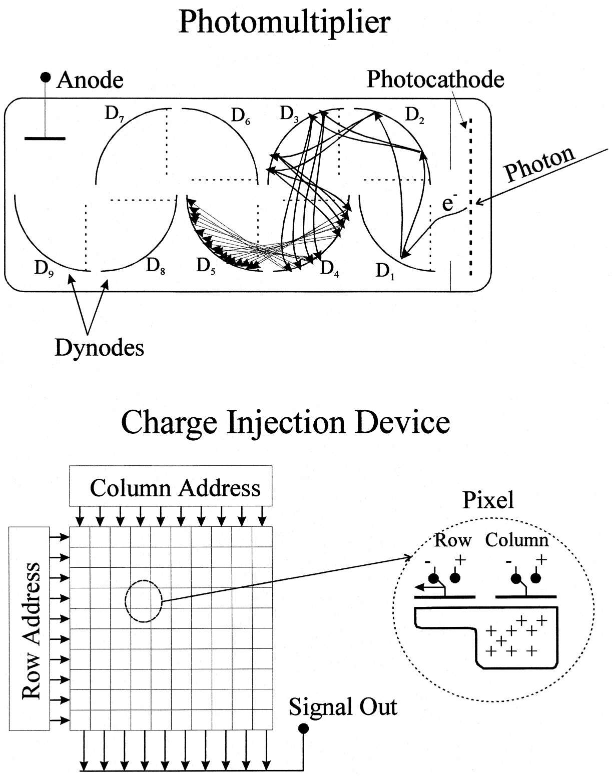


FIGURE 4 Examples of single-channel (photomultiplier) and two-dimensional array (charge injection device) detectors for spectroscopic applications.

III. ATOMIC EMISSION SPECTROMETRY

A. Theoretical Background

In atomic emission spectrometry, the intensity of emission from the analyte atoms depends on the number of atoms that are in the excited state. If the temperature of the atom cell is increased, then more atoms undergo energy exchange collisions with the surrounding hot gases. More atoms then possess enough energy to be excited to an upper state and drop back down to the ground state by loss of a photon. The expression that relates the temperature of the atom cell to the number of atoms in the excited state is the Boltzmann expression. If n_0 is the number of atoms in any given state with energy E_0, then the number n^* in an excited state with energy E_q, is given by

$$n^*/n_0 = g_q/g_0 e^{-(E_q - E_0)/kT} \quad (1)$$

where k is the Boltzmann constant, and T is the absolute temperature. The constants g_q and g_0 are the statistical weights (degeneracy), which account for the fact that there are often several states with the same energy E_q and E_0, respectively.

The Boltzmann expression predicts the number of atoms in the excited state, relative to the number of atoms in the ground state. For example, considering the atomic emission from Cs (852 nm; $E_q - E_0 = 1.46$ eV) and Zn (213.8 nm; $E_q - E_0 = 5.8$ eV) at 4000 K, the n^*/n_0 ratios for these elements are 2.98×10^{-2} ($g_q/g_0 = 2$) and 1.48×10^{-6} ($g_q/g_0 = 3$), respectively. Another way to use the above equation is to generate a plot for the Boltzmann factor ($e^{-(E_q - E_0)/kT}$) as a function of the excitation temperature, as shown in Fig. 5. For a given transition (ΔE), the Boltzmann factor becomes dependent only on the temperature; as a result, higher temperatures within the atom cell will generate a larger population of the excited species, which ultimately leads to higher emission intensities.

Flames have temperatures in the 2000–3000 K range, while atmospheric pressure plasmas are in the 4000–10,000 K range. In the air–acetylene flame, which has a temperature of about 2540 K, it is possible to see relatively large amounts of light emitted by sodium, cesium, calcium, and other metals, but it is very difficult to see zinc emission in the same flame. If a nitrous oxide–acetylene flame is used, with its typical temperature of 3150 K, larger signals from zinc and many other metals can be seen. In general, flames are not a convenient source for atomic emission measurements in light of the fact that in plasmas, most elements can be determined with high sensitivity by AES.

The primary variable that determines the temperature necessary to excite strong atomic emission signals from a particular element is the energy difference ($E_q - E_0$). The

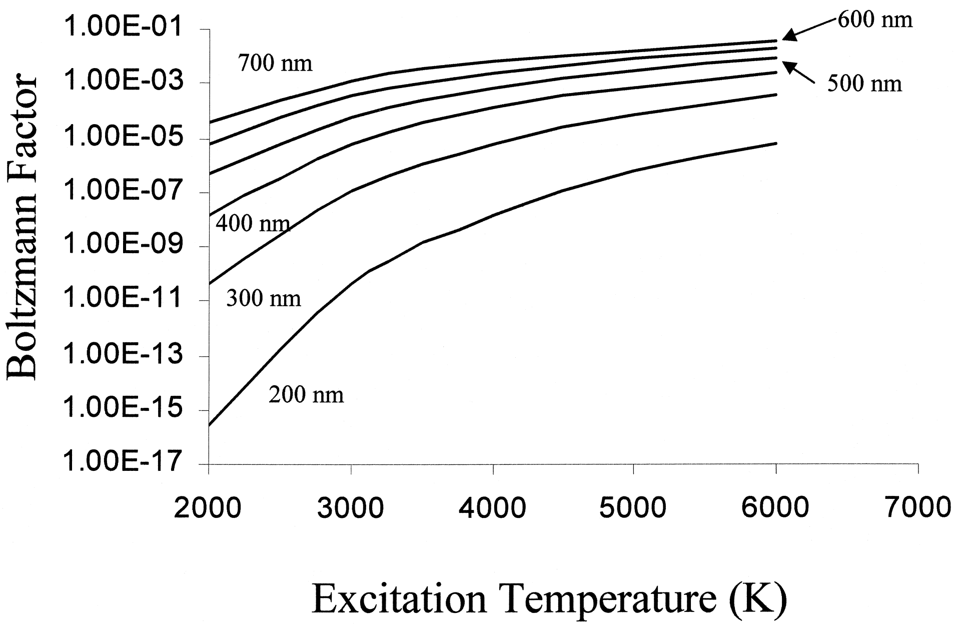


FIGURE 5 The Boltzmann factor as a function of the excitation temperature. Each curve represents a particular ΔE in the ultraviolet or visible region of the electromagnetic radiation.

larger this difference, the more energy required to cause the element to be excited to the higher state. In general, elements with analytical lines (usually resonance lines) in the blue or ultraviolet do not give a strong atomic emission signal in a flame. These trends can be seen in Fig. 5, where the analytes with short analytical wavelengths have a correspondingly low Boltzmann factor.

In addition to the Boltzmann equation, the following expression is used to account for other variables that affect the atomic emission signal at low concentrations:

$$\Phi = Ah\nu_0 n(g_q/g_0)aL(\Omega/4\pi)e^{-(E_q-E_0)} \qquad (2)$$

where Φ is the radiant power (in watts) emanating from the atom cell, and n is the number of atoms in the atom cell per cm^3. A is the Einstein transition probability, which indicates the probability that an excited atom, per unit time, will fall into a lower level by spontaneous emission of a photon. This is a rate constant for the first-order decay of the atom from the excited state to the lower state. In the above equation, a is the area of the atom cell observed by the detector, L is the thickness of the atom cell, and Ω is the solid angle (in steradians) over which the emission is being observed. There are other losses that are often difficult to quantify, such as light losses in the optical components. In principle, if there are n^* atoms per cm^3 in the excited state, then n^*A photons with energy $h\nu_0$ will be emitted per unit time and per cm^3. In the case of sodium, which has two excited states of similar energy (589.0 and 589.6 nm) and similar A values, the relative emission intensities are the same as the ratio of their statistical weights (1 : 2).

During the past two decades, plasma spectroscopy has dominated the analytical market for atomic emission spectrometry. Consequently, we will not discuss the use of flames for atomic emission, as the industry, in large, has moved toward plasmas for this application. On some occasions, flame emission is used for analysis of sodium and potassium. Nonetheless, these elements can be easily determined by inductively coupled plasma spectrometry.

B. Inductively Coupled Plasma Spectrometry

1. Instrumentation

A diagram of the inductively coupled plasma (ICP) device is shown in Fig. 6. The effective temperature of plasma is two to three times higher than those observed in flames (in the 4000–10,000 K region). Passing argon gas through a set of concentric quartz tubes within a strong radio frequency (rf) field generates the plasma. Radio-frequency energy at about 27 MHz and 2 kW of power is coupled to the water-cooled induction coils at the top of the quartz plasma tube. This energy is sufficient to accelerate ionized particles into many collisions which causes further ionization and emission of significant background radiation from the argon gas. The plasma is ignited by seeding with electrons generated by an electrical discharge. After the onset, the plasma becomes a self-propagating and -sustaining process. The tangentially introduced argon gas (Fig. 6) spins the plasma into a torroidal shape at the point where the sample is introduced. The sample is injected

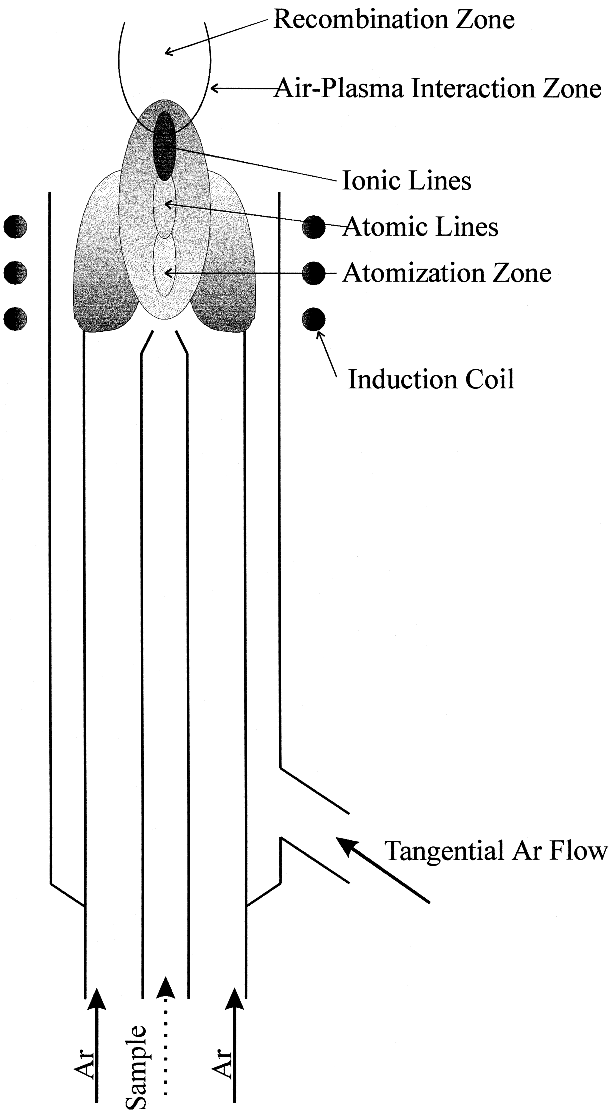


FIGURE 6 Schematic representation of an inductively coupled plasma torch during operation.

into the center of the plasma where excitation is more efficient. As shown in Fig. 6, the plasma may be divided into observation zones where atomic or ionic species are more abundant.

The sample is introduced into the plasma commonly through a nebulizer. The function of the nebulizer is to convert the sample solution into a fine aerosol spray. The nebulizers are often coupled to a spray chamber where large droplets are preferentially removed from the aerosol stream. Discrimination against larger droplets facilitates more stable plasmas and a better measurement precision. The most frequently used nebulizers are depicted in Fig. 7. Cross-flow nebulizers are robust aerosol generators suitable for most applications. The high-efficiency nebulizer (HEN) is ideally suited for applications requiring efficient nebulization at low solution-flow rates. In general, while the ultrasonic nebulizer is more expensive than HEN and cross-flow systems, it does provide for a better analytical detection limit.

2. Instrumentation for Multi-Element Analysis

One of the primary advantages of an ICP is that it provides good sensitivity for the determination of a wide range of elements without significant chemical interferences. With the use of any multichannel detection system, one can detect the emission coming from many elements simultaneously. The spectrometers shown in Fig. 3 are capable

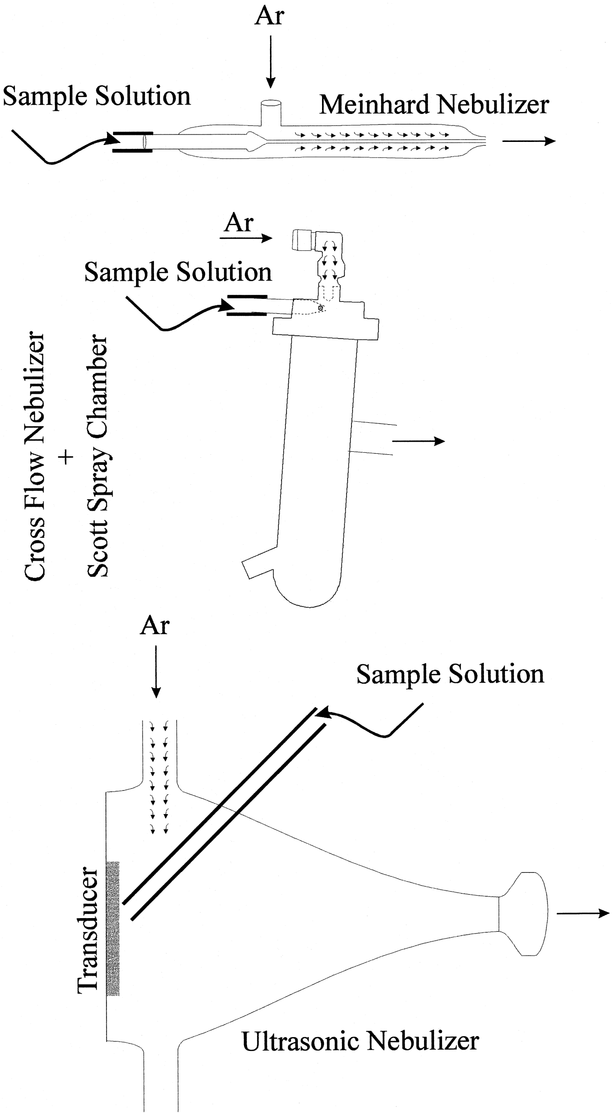


FIGURE 7 Three popular nebulizers used for introduction of solutions into an atom cell.

of simultaneous interrogation of many spectral lines. The electronic data-handling system is then able to gather data from a large number of elements simultaneously. Thirty elements are typical, and as many as 60 elements are practical for an optimized analytical run.

3. Sensitivity

Atomic emission signals in the ICP are much larger than those in the flame for nearly all elements and can be obtained for a wider range of elements. The high-temperature, inert-argon environment of the ICP leads to a more complete atomization and efficient excitation of analytes, resulting in larger signals.

The copious energy in the ICP causes some atoms to become completely ionized. For analytes with extensive ionization, the atomic emission signal is minute, but the ionic emission signal is quite substantial. The concept of an ionic emission signal is exactly the same as that of atomic emission. An ion has a different set of energy levels from the equivalent neutral atom (i.e., ionic emission signals occur at different wavelengths from the atomic emission signals). This duality of ICP spectrometry has led to the use of the name optical emission spectrometry (OES) for the ICP emission technique rather than atomic emission spectrometry. Signals from ion transitions are used just as routinely as atom lines for the determination of the concentration of an analyte. Furthermore, the ions from an ICP may be extracted into a mass analyzer for elemental mass spectrometry (discussed later).

4. Chemical Interference Effects

One of the most important advantages of the ICP over other atomic emission sources is the near absence of chemical interferences. The argon plasma is inert and the temperature is so high that nearly all compounds are broken down efficiently. Virtually all the analyte is released for excitation and subsequent atomic emission.

5. Spectral Interferences

Because of the high excitation temperatures attained in ICP, the most serious disadvantage of this plasma is the relatively large number of spectral interferences. Spectral overlap becomes very likely in complex samples containing many elements over a wide range of concentrations. The potential spectral overlaps are often remedied with the use of high-resolution monochromators (polychromators).

The atomic emission signal from a particular element is not composed strictly of one wavelength but of a very narrow range of wavelengths (approximated by Fig. 8). The width of the spectral line at half-height (the half-width) is about 5 pm, but it can vary by a factor of ten or more depending upon many fundamental and experimental parameters (e.g., temperature and pressure).

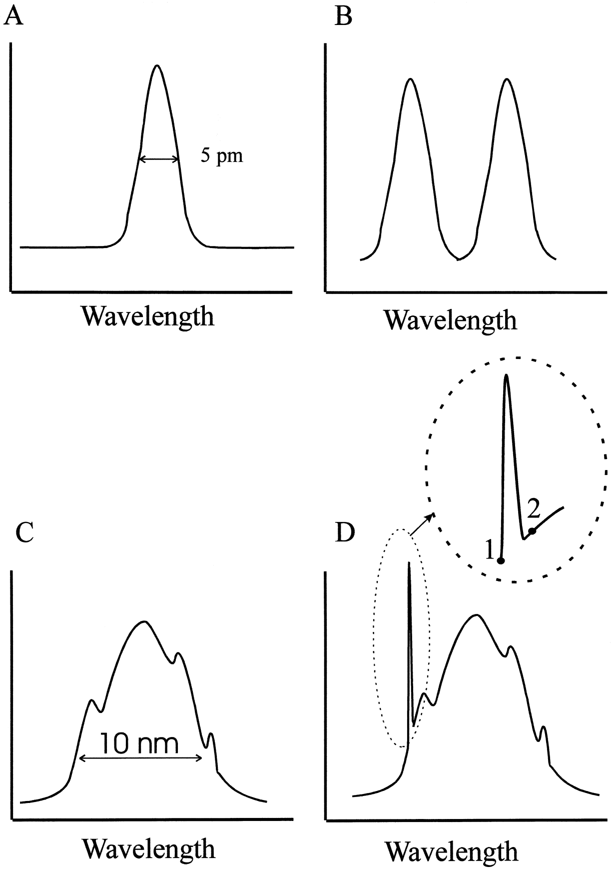


FIGURE 8 Atomic emission profile and potential spectral interferences. (A) A typical atomic emission line profile intensity as a function of wavelength (note the narrow width of the emission line). (B) Two independent atomic emission profiles emitted from two different elements that can be resolved with the use of a high-resolution monochromator. (C) A typical molecular emission band as a function of wavelength (note the emission coverage over a wide wavelength range as compared to atomic emission). (D) Atomic emission superimposed on a concurrent molecular emission (points 1 and 2 represent the background correction regions for the atomic emission signal).

Some analytes emit lines that are very close to the lines of other analytes (Fig. 8B). For example, iron emits radiation at a multitude of lines, any of which might be at almost the same wavelength as that emitted by another metal. If the slits on the monochromator are narrow enough and if the dispersion of the monochromator is large enough, spectral lines that are very close to each other will be well separated. The resolution of a monochromator is said to be high if it is able to separate close spectral lines. The highest possible resolution is usually required for

TABLE I Elements with Pronounced Spectral Interference for Determination of Be Using 313.042 nm Line

Element	Observed changes on Be emission[a] (%)
Cr	−120
Ta	−20
U	−15
Ag	0
Gd	+30
Th	+75
Ti	+160

[a] Positive changes in Be emission intensities are due to direct line or wing overlaps. Negative deviations are due to interference at a background correction point.

atomic emission spectrometry, but most of the monochromators that are commonly used in emission spectrometry cannot resolve two lines that are closer than about 3 pm apart.

If two spectral lines from two metals are not satisfactorily separated, then a spectral interference occurs. Spectral interferences can still occur even with a large, high-resolution monochromator. The resulting measured signal not only represents the element of interest, but also a component of the signal resulting from the interfering element. In general, the signal is larger than it should be, which results in an inaccurate measure of the element concentration. As an example, a list of elements with spectral interference for determination of Be is given in Table I.

The monochromators that are used for ICP optical emission spectrometry (ICP–OES) often include automatic scanning capability to measure the background at each side of the analytical atom or ion line. The primary approach to discriminating against spectral line interferences is through the use of a high-resolution monochromator. If this fails, then one recourse is to find another atomic or ionic emission line that is sufficiently sensitive but which is not affected by a spectral interference. Examples of acceptable alternative emission lines for some elements are presented in Table II. Alternatively, it is possible to remove the element that is causing the spectral interference by use of solvent extraction, or another chromatographic technique, prior to sample introduction.

The atomic emission from inductively coupled plasma frequently resides on a substantial background emission. The majority of background continuum emission is due to *bremsstrahlung* radiation (deceleration of fast-moving electrons) or an electron-ion recombination process. Another cause of the background radiation is the molecular band emission from the OH species originating from the aqueous samples (Fig. 8C). As shown in Fig. 8D, the scanning capability of monochromators may be used to correct for these types of interferences. As seen in Table I, the presence of some elements (e.g., presence of Cr, Ta, and U while measuring the emission from the Be 313.042 line) may cause an over-correction of the background emission.

Spectral interferences can be corrected if the magnitude of the interference is known as a function of the concentration of the interfering element. A correction factor may be calculated and ratioed to the signal size at the analyte wavelength (concentration ratio method). This is best accomplished by making simultaneous measurements of both the signal at the analyte wavelength and the concentration of the interfering element at another wavelength. The main requirement is that the concentration of the interfering element can be measured at the other analytical wavelengths without any spectral or other interferences. Hence, polychromators are ideally suited for this procedure. It is also assumed that both the interferent signal, which causes the spectral interference at the analyte wavelength, and the interferent signal at the second wavelength behave in the same way in the plasma. This turns out to be an acceptable method of correction for spectral interferences due to the lack of chemical interferences within the plasma.

An internal standard is frequently used to correct for drift in the signal size as a function of time. An internal standard is an element of known concentration in a sample that is present in all examples of the sample. The instrument monitors its signal, often simultaneously with the measurements described, and if it drifts in size an appropriate correction is applied to the analyte signal to compensate for the drift. This helps to ensure that the instrument remains accurately calibrated for the duration of the analytical run.

TABLE II Analytical Wavelengths Used in ICP-AES Analysis

Analyte	Primary analytical wavelength (nm)	Alternate analytical wavelength (nm)	Decreased sensitivity factor
Ag	328.068	338.289	2.0
Au	242.795	267.595	1.8
Be	313.042	234.861	1.2
Cd	214.438	228.802	1.1
Cu	324.754	224.700	1.4
Ga	294.364	417.206	1.4
In	230.606	325.609	1.9
Ni	221.647	232.003	1.3
Pb	220.353	216.999	2.1
Zn	213.856	202.548	2.2

C. Electrical Discharges for Optical Emission Spectrometry

1. dc and ac Arcs and Sparks

Arcs and sparks were the dominating emission techniques nearly four decades ago and in most arenas they have been replaced by ICP emission spectrometry. Nonetheless, arcs and sparks have escaped complete extinction due to their versatility for analysis of solid samples. These techniques are still in use in many foundry-based industries and nuclear fuel-fabrication laboratories. These types of electrical discharge are comprised of two electrodes, across which an electric current is passed. In the case of the dc and ac arcs, a low voltage of 10–50 V is used and a current of 1–35 A flows between the electrodes. The ac arc is a series of separate discharges that occur once during each half-period of the power cycle. The ac spark is based on the discharge of a capacitor that has been charged to 1–30 kV. The spark occurs 120–1800 times per second. The temperatures of these discharges are in the same range as the ICP; hence, many species can be excited, and quantitative and qualitative analyses can be obtained.

The sample is typically placed in a cup in the bottom electrode, so the discharge occurs between the anode electrode and the sample (cathode). The sample can be a conductive solid or it may be crushed and mixed with a conducting material such as powdered graphite.

The instrumentation for these discharges is essentially the same as those used for all emission experiments. The detection system after the monochromator (also known as a *spectrograph*) can be the multiple photomultiplier arrangement, the older photographic plate arrangement, or a charge injection device. The intensity of the image of each line is proportional to the amount of light emitted from the discharge at each wavelength, and the concentration can thus be interpolated by use of standards of known concentration.

Electrical discharges are affected by a number of serious matrix interferences associated with the way that the sample is vaporized into the discharge as well as various chemical and physical interactions within the discharges. Spectral interferences are as serious as for ICP–OES. These interferences can be mitigated, in part, by the use of internal standards and the concentration ratio method. However, these approaches are not as successful as with the ICP because the latter has fewer interferences and better precision. Internal standards have to be chosen with great care because they must behave in the same way as the analyte. This is not trivial because the physicochemical interference problems that occur in these atom cells vary greatly from element to element and sample to sample.

2. dc Plasma

The dc plasma (DCP) technique is similar to ICP and less like the electrical discharges described above. Figure 9 shows the basic arrangement of the three-electrode DCP system. The plasma is formed by the discharge of dc current at low voltage across graphite electrodes bathed in argon. The main difference between the DCP and the dc arc is that the sample is introduced in liquid form. The sample is aspirated and converted into small droplets (using a nebulizer), in a way similar to the ICP systems. The premixed droplets and argon are then allowed to flow into the DCP observation zone. The sensitivity of the plasma for the determination of metals in samples is inferior to that of ICP. There are minimal chemical interferences, and spectral interferences are as serious as with ICP. DCP is also affected by easily ionized element (EIE) interference, which cause a 30–80% enhancement of the signals of some elements. DCP instrumentation is, in principle, similar to that of the ICP, because multi-element analyses are possible by use of a spectrometer that has multiple photomultiplier tubes.

3. Glow Discharge Plasma

A direct current glow discharge (GD) plasma is formed in a low-pressure, inert buffer gas (e.g., 0.1–10 torr of Ar). The sample is composed of conductive material, which becomes a part of the electrical circuit (cathode). The discharge is typically sustained at several hundred volts at a current of a few mA. GD plasmas have a high charge density and large electric field gradients (kV/mm) near the cathode surface. Initially, the argon buffer gas becomes ionized and is accelerated toward the cathode due to a net electric field. As the fast moving Ar impinges on the cathode surface (sample), the sputtering process removes a few layers of material. The sputtered material is subsequently excited and ionized through a series of complex gas-phase processes. These include electron impact ($A + e^- \rightarrow A^*, A^+, A^{*+}$), Penning ionization by the metastable species of the buffer gas ($A + Ar^m \rightarrow A^{+*} + Ar$), radiative recombination ($A^+ + e^- \rightarrow A^* + h\nu$), and radiation trapping ($A + h\nu \rightarrow A^*$).

Two types of glow discharge sources are illustrated in Fig. 10. The hollow cathode GD is often used as a line source for many spectroscopic applications (such as atomic absorption, which is discussed in the next section). A potential of up to 400 V is placed between the anode and cathode of the lamp to initiate the plasma. During operation, an electrical current between 4 and 40 mA at 150–350 V sustains the plasma in a low pressure of an

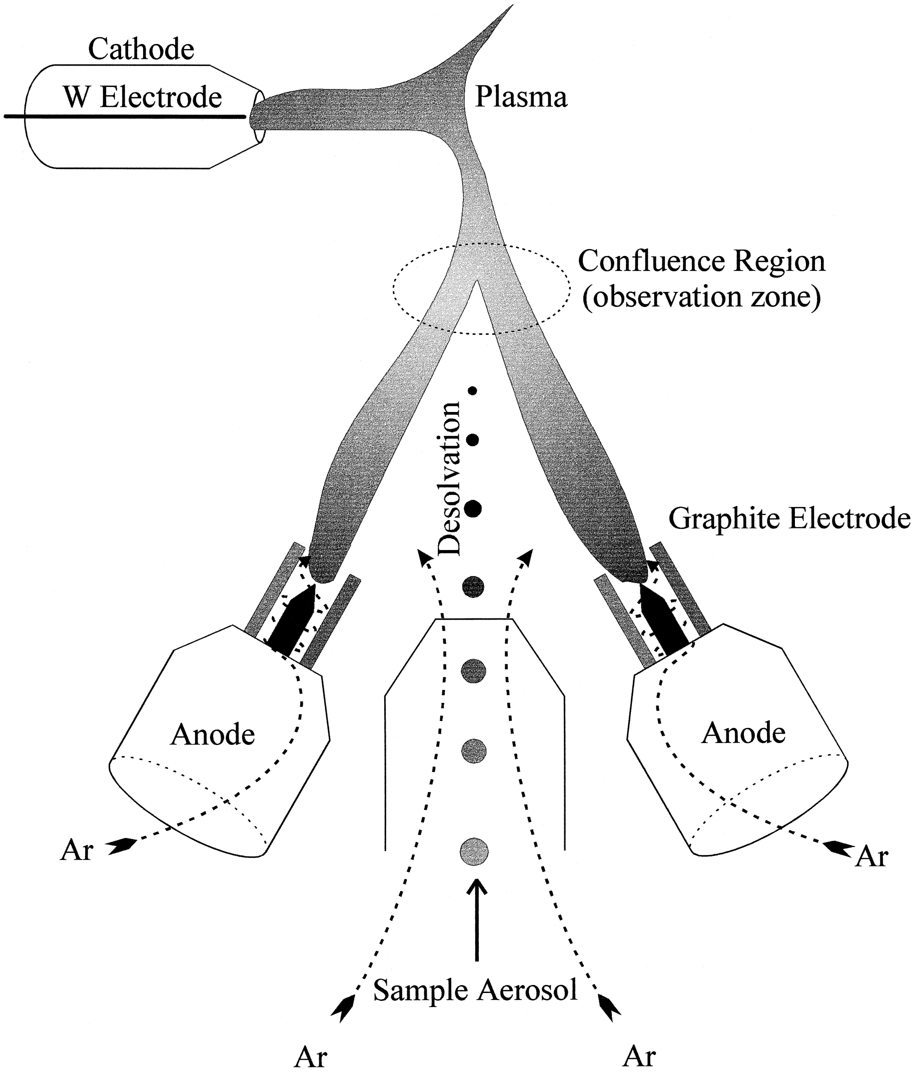


FIGURE 9 A direct current plasma emission source.

inert gas (argon or neon). The cathode is made of the element that is to be determined in a sample. Occasionally, these plasma sources may be used for determination of elements in a solution matrix. In these instances, a few microliters of a solution are deposited and dried onto the cathode surface. The chamber is then evacuated and backfilled with a small amount of Ar gas (1–10 torr). For dried solutions, the analyte emission is a transient signal that is integrated over time and compared to integrated intensities for a standard solution.

The Grimm source is the most popular GD design for spectrochemical determination of analytes in conductive solids (Fig. 10). The plasma is a restricted discharge on the cathode yielding uniform sputtering and well-defined craters. The calibration curves are always established with known analyte concentrations in similar matrices.

Another popular approach to glow discharge spectroscopy is to use rf power instead of traditional dc power sources. The main advantage of rf-GD is its ability to sputter nonconductive samples, hence elemental analysis for polymers and ceramics becomes a matter of simple solids analysis.

4. Other Plasmas

Microwave-induced plasmas (MIP) are generated using the microwave region of electromagnetic radiation (1–3 GHz). The microwave is generated by a magnetron source, and the energy is coupled into an inert gas (e.g., He) containing the sample, using a waveguide and a cavity. About 50–200 W of power are needed to sustain the plasma. MIP is an ideal source for analysis of nonmetals

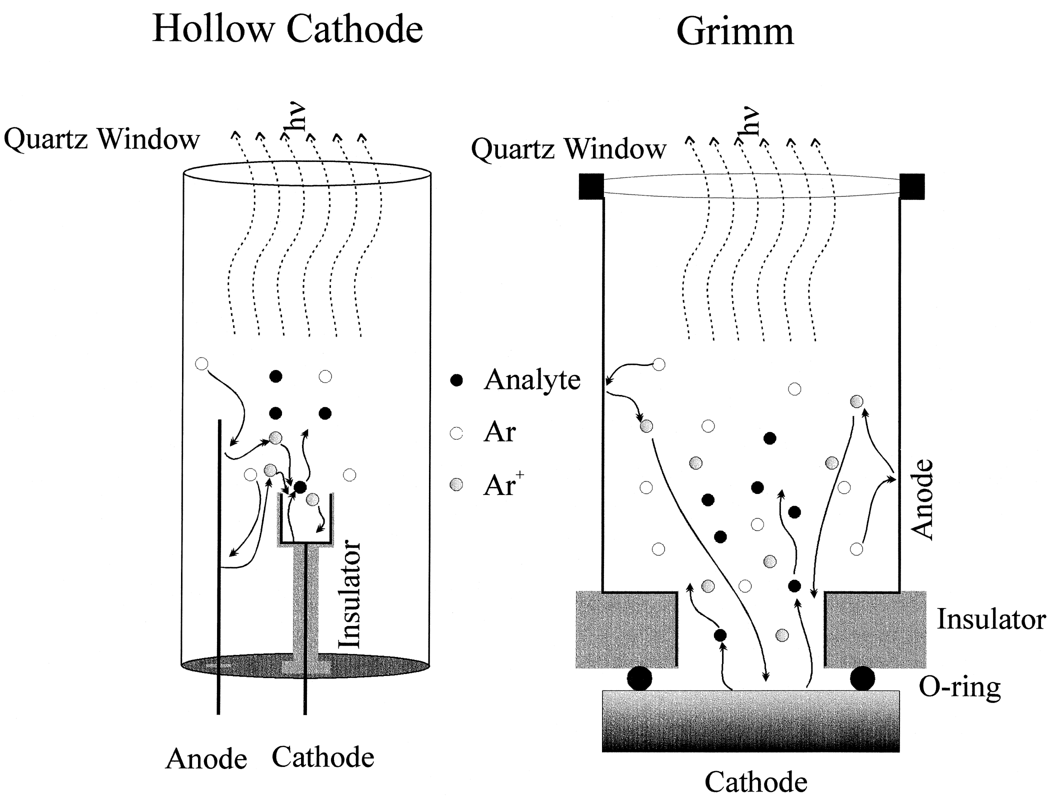


FIGURE 10 Two different glow discharge cells based on the hollow cathode and the Grimm-type design.

but it is not the most robust source for analysis of aqueous solutions. MIP has been successfully, applied to gas chromatography applications as an atomic emission source.

Laser-induced plasma spectroscopy (LIPS), which is also known as laser-induced breakdown spectroscopy (LIBS), is a versatile tool for elemental analysis. A pulsed laser beam (lasting a few nanoseconds) is focused onto a target at power densities in excess of 100 MW/cm^2. The plasma is formed when the irradiance of the laser beam is high enough to generate an electric field at the focal volume that exceeds the breakdown threshold of the target medium. Laser plasmas with well-defined spatial resolution can be formed in almost any medium. Good analytical results may be obtained for LIBS instruments equipped with time-resolved detection technology. Laser-induced plasmas can be optimized to yield a detection limit of tens of μg/g for certain elements in ideal matrices.

IV. ATOMIC ABSORPTION SPECTROMETRY

A. Theoretical Background

Figure 1B illustrates the basic principle of atomic absorption spectrometry (AAS). The amount of light absorbed from the light source is related to the transmittance T, where:

$$T = I/I_0 \tag{3}$$

I_o and I are the initial intensity of the source and the source intensity after the atom cell, respectively.

The path length b in the atom cell (Fig. 1F) needs to be relatively long to maximize the amount of light absorbed by the analyte. The amount of light absorbed depends on the fundamental constant k', the *atomic absorption coefficient*. These parameters are related to transmittance in the following manner:

$$I/I_0 = e^{-k'b} \tag{4}$$

The absorbance A is the parameter that is usually determined with an AAS instrument, where:

$$A = -\log I/I_0 = \log 1/T = k'b \log e = 0.434k'b \tag{5}$$

The constant k' prime is proportional to a number of parameters, including the number of atoms per unit volume, the Einstein transition probability for the absorption process, and the energy difference (Fig. 1) between levels 0 and 1. The Beer-Lambert law combines these constants into one constant a to yield the following equation:

$$A = abc \tag{6}$$

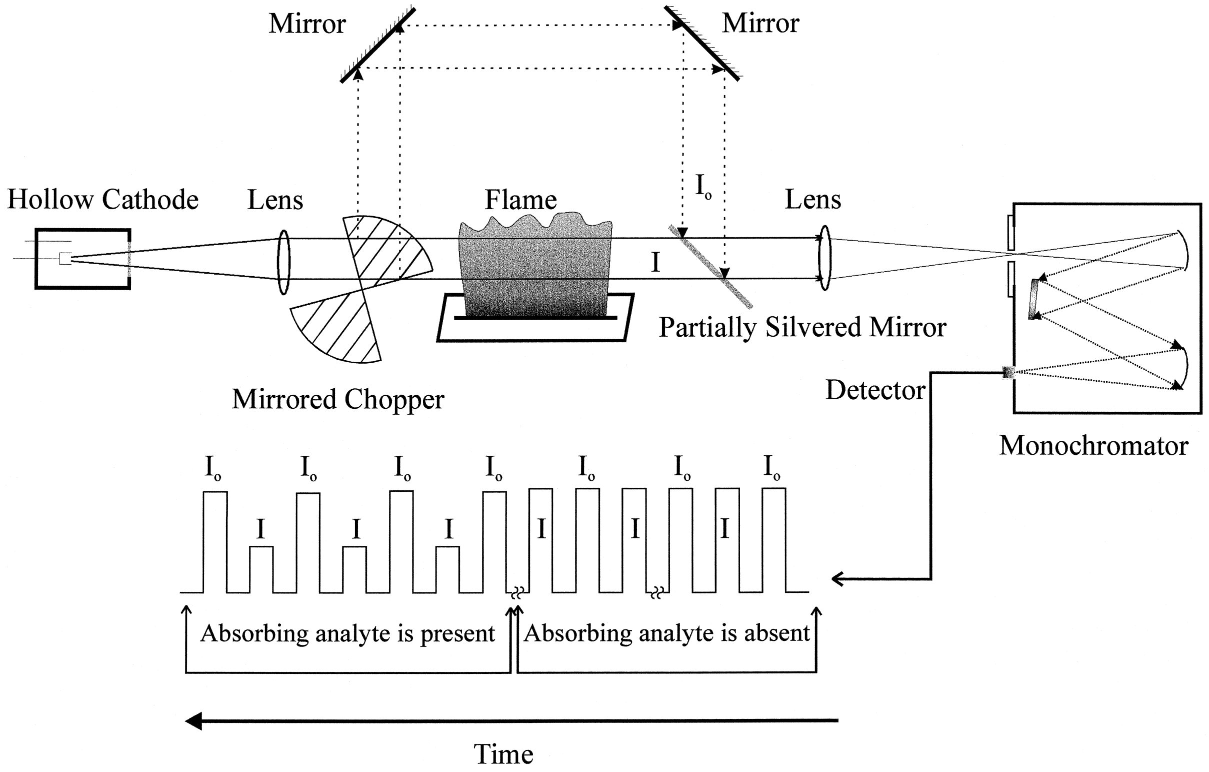


FIGURE 11 A dual-beam flame atomic absorption instrument with the corresponding timing diagram.

where a is a constant called the *absorptivity* and c is the concentration of the analyte in solution.

The signal in an atomic absorption instrument is the measured reduction in the light intensity from I_0 to I. In modern instruments, the logarithm is automatically calculated so that the output of the instrument is the absorbance. Absorbance is a unitless number that varies typically between 0.001 and 2, but the number is usually referred to as *absorbance units*. The AAS calibration curve is a plot of A versus c with a slope of ab. It is best to work at concentrations that give signals in the middle of the absorbance range.

B. Flame Atomic Absorption Spectrometry

Flame AAS may be used to determine the concentration of most metals by using an acetylene-based flame. An instrumental arrangement for a flame atomic absorption spectrometer is presented in Fig. 11. There are virtually no spectral interferences that affect AAS. Nonetheless, the main limit to analysis by AAS is due to physico-chemical interferences. The proliferation of ICP instruments has significantly reduced the overall application of flame AAS in most industrial applications. More than likely, the use of flame atomic absorption will dwindle to only a few specialized applications (similar to flame emission instruments).

1. Instrumentation

The hollow cathode lamp: The most important aspect of AAS instrumentation is the light source, the most common of which is the hollow cathode lamp (HCL). Figure 10 illustrates the basic construction details of these lamps. Other line sources such as electrode-less discharge lamps are frequently used for determination of nonmetals. In addition to dc operation, hollow cathode lamps may be operated by supplying pulsed current (at about 50–200 Hz) to increase the average amount of light that can be obtained from the lamp. *Atomic absorption profile* is the term used to describe the range of wavelengths over which it is possible for an atom to absorb radiation. The amount of light absorbed is proportional to the atomic absorption coefficient k' (Fig. 12A). In most atmospheric pressure atom cells, the profile has a typical width at half-maximum of 5 pm or greater and is a function of temperature and the concentrations of major species in the atom cell. The primary assumptions of the Beer–Lambert law is that the half-width of the light source is narrower than the half-width of the atomic absorption profile. The HCL is almost an ideal source for AAS because of its narrow spectral line profile (half-width = 2 pm). Figure 12A shows a representation of the overlap of the light source emission profile with the atomic absorption profile. If the light source spectral line width is significantly broadened (i.e., it is comparable to

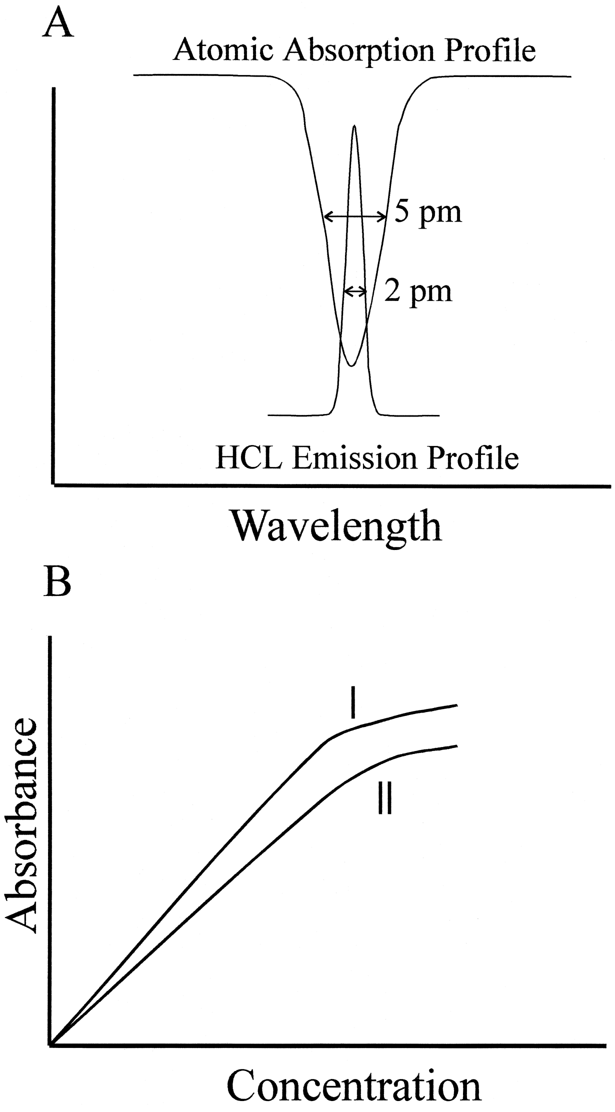


FIGURE 12 Influence of line shape on calibration parameters for atomic absorption spectrometry. (A) Comparison of atomic line widths for a hollow cathode lamp versus the atomic absorption line width observed in atmospheric pressure atom cells. (B) The slope and linear dynamic range of the calibration changes from "I" to "II" as the emission line width of the hollow cathode lamp becomes broader.

the absorption profile), then the slope of the calibration curve changes, as illustrated in Fig. 12B. This change in the slope results in a loss in sensitivity. Such broadening of the light source emission profile occurs in practice as the HCL ages or if it is run at operating currents that are too high. Regardless of the lamp output profile, AAS tends to have calibration curves that have a limited *linear dynamic range* (the straight part of the calibration curve) relative to the other atomic spectrometric techniques (AES and AFS). The AAS linear dynamic range is typically about two orders of magnitude as a function of concentration. In contrast, ICP, AES, and AFS have linear dynamic ranges of 3 to 6 orders of magnitude.

The flame: For AAS measurements it is necessary to break down compounds in the flame into their constituent elements. Air–acetylene and nitrous oxide–acetylene are the most frequently used flame gases. Most elements can be determined by AAS in the air–acetylene flame. The nitrous oxide–acetylene flame is necessary for 5–15 elements whose compounds do not break down sufficiently in the air–acetylene flame. Aluminum, for example, does not produce a signal in the air–acetylene flame, while calcium gives signals in both flames but a larger signal in the nitrous-oxide flame. The burner used is usually a slotted burner (Fig. 11), which allows for the maximum signal due to the long absorption path length.

Modulation techniques to discriminate against background signals: The experimental arrangement for optical modulation is shown in Fig. 11. As depicted in the figure, the hollow cathode lamp is placed on the optical axis of the monochromator. In this arrangement, the incident HCL radiation is split into two parts, one passing through the flame and the other passing around the flame. A mirrored optical chopper is used in front of the HCL so that the light output is periodically interrupted and rerouted around the flame. This is called *intensity modulation* of the light source output and is used to allow discrimination against background signals. The two main background signals that affect AAS (and AFS) measurements are the background light emitted from the atom cell and the atomic emission signals from the elements in the sample matrix. Atom cells always have some inherent background luminosity. While the background in the flame is due to the emissions that result from the chemical reactions within the flame, graphite furnace atomizers (see next section) emit blackbody radiation. All atom cells are hot enough to excite some atomic emission signals. Consequently, absorption (and fluorescence) signals are always accompanied by atomic emission signals from most of the matrix elements. All background signals must be subtracted to avoid the spectral interferences associated with the atomic absorption measurements. Figure 11 helps demonstrate how modulation subtracts out these background signals and is an illustration of the temporal behavior of the signals that are observed by the detection system at a particular wavelength. The flame background emission and the matrix atomic emission signals are emitted continuously from the atom cell while the sample is being introduced and are represented in Fig. 11 as the baseline. The hollow cathode lamp emission is also observed by the detection system, but the chopper periodically switches off its signal. While the lamp is blocked, a measure of the total background can be made and then subtracted from the HCL signal plus background signal that is obtained when the HCL is on.

Hence, all backgrounds, including any atomic emission spectral interferences, are subtracted out automatically.

The monochromator and detection system: The monochromator for AAS does not need to have the high resolution necessary for AES, because the selectivity of AAS depends on the line width of the light source as discussed above. The role of the monochromator is to reject the majority of the background radiation as well as selecting only the pertinent HCL emission line for AAS measurements, thus monochromators for AAS have a moderate resolution of 0.02–2.0 nm. The detection system is based on a photomultiplier tube, and the readout electronics are similar to single-element AES instruments.

2. Interference Effects

Physico-chemical interference effects: The flame atomization process is affected by chemical interferences that prevent facile formation of neutral metal species of M. For example, oxides of many metals are likely to form in the flame environment. Oxygen is abundant in any flame as a consequence of the oxidant (air, oxygen, or nitrous oxide) used in the combustion process. The main concern is whether or not there is enough energy in the flame to break the bond between the metal and the oxygen to release the metal. If the flame does not have enough energy, most of the metal oxides remain intact and very little free metal is available. Phosphates also cause chemical interference due to formation of a stable compound in the air-acetylene flame that breaks down slowly to release the metal atom. This reduces the expected signal size significantly. In striving to discriminate against chemical interferences, the main aim is to obtain some assurance that the signal measured by the detection system does indeed quantitatively represent the concentration of the analyte in any particular matrix. We try to calibrate the instrument with a known concentration of the analyte in a deionized water matrix and then compare directly the signal from a real sample with the calibration curve. The real sample components in the matrix, such as phosphate in a biological matrix during the determination of calcium, cause the signal to be smaller (a depression) or sometimes larger than the signal obtained from the same concentration of analyte in the calibration solution. One way to circumvent chemical interferences is to use releasing agents. Lanthanum chloride, which is usually added (1% weight/volume) to all sample solutions, is a good example of a releasing agent. The lanthanum oxide and phosphate compounds are often more stable than the same compounds of other metals. In the flame, the lanthanum oxide or phosphate forms in preference to the analyte oxide or phosphate and thus releases the analyte atom. Other releasing agents that have been found to work in many situations are strontium chloride and complexing agents such as EDTA (ethylenediaminetetraacetic acid). Releasing agents increase the magnitude of the signal in real samples. They can also change the magnitude of the signal from the standard solutions.

Spectral interferences: The elegance of atomic absorption lies in the very high selectivity of the technique. The HCL only emits light that is characteristic of the elements in the cathode. While in the atom cell there may be many elements present, only one element that corresponds to the cathode material absorbs the light, because only the light characteristic of that element is emitted by the light source and reaches the detector. The HCL acts like a probe to determine the concentration of only one metal. This means that virtually no spectral line interferences affect AAS. The technique is often much simpler to use than AES, and the instrumentation can be relatively inexpensive while providing high sensitivity and selectivity.

C. Graphite Furnace Atomic Absorption Spectrometry

Graphite furnace AAS, also known as electrothermal atomic absorption spectrometry (ETV–AAS), is one of the most sensitive techniques available for routine elemental analysis. This technique is capable of the determination of picogram (10^{-12} g) amounts in a few microliters of sample.

1. Instrumentation

The arrangement for this technique is similar to that for flame AAS, except that a graphite furnace replaces the flame. The atomizer consists of a graphite tube about 3 cm long, with a 6-mm internal diameter and wall thickness of about 1 mm. Two different graphite tube designs are shown in Fig. 13. The tube is fixed between two electrodes and is subjected to a low-voltage (up to 12 V), high electrical current. The power supply can be programmed to heat the furnace to several pre-selected temperatures. The furnace has an inert gas passing through and around it to minimize the ingress of oxygen from the air, which would lead to rapid oxidative degradation of the furnace tube. Five to 100 μl of the sample solution are placed in the furnace and dried at 100°C. The sample is then pyrolyzed to break down or volatilize the matrix component. The pyrolysis temperature is set high enough without unintentionally vaporizing the analyte (about 400–1200°C). After pyrolysis, the furnace is rapidly heated to the atomization temperature to volatilize the analyte into the atom cell. The atomization temperature can be anywhere in the range of a few hundred to about 2700°C. The analyte then absorbs light from the hollow cathode lamp to give the atomic absorption. This signal is a transient (Fig. 13E), because the

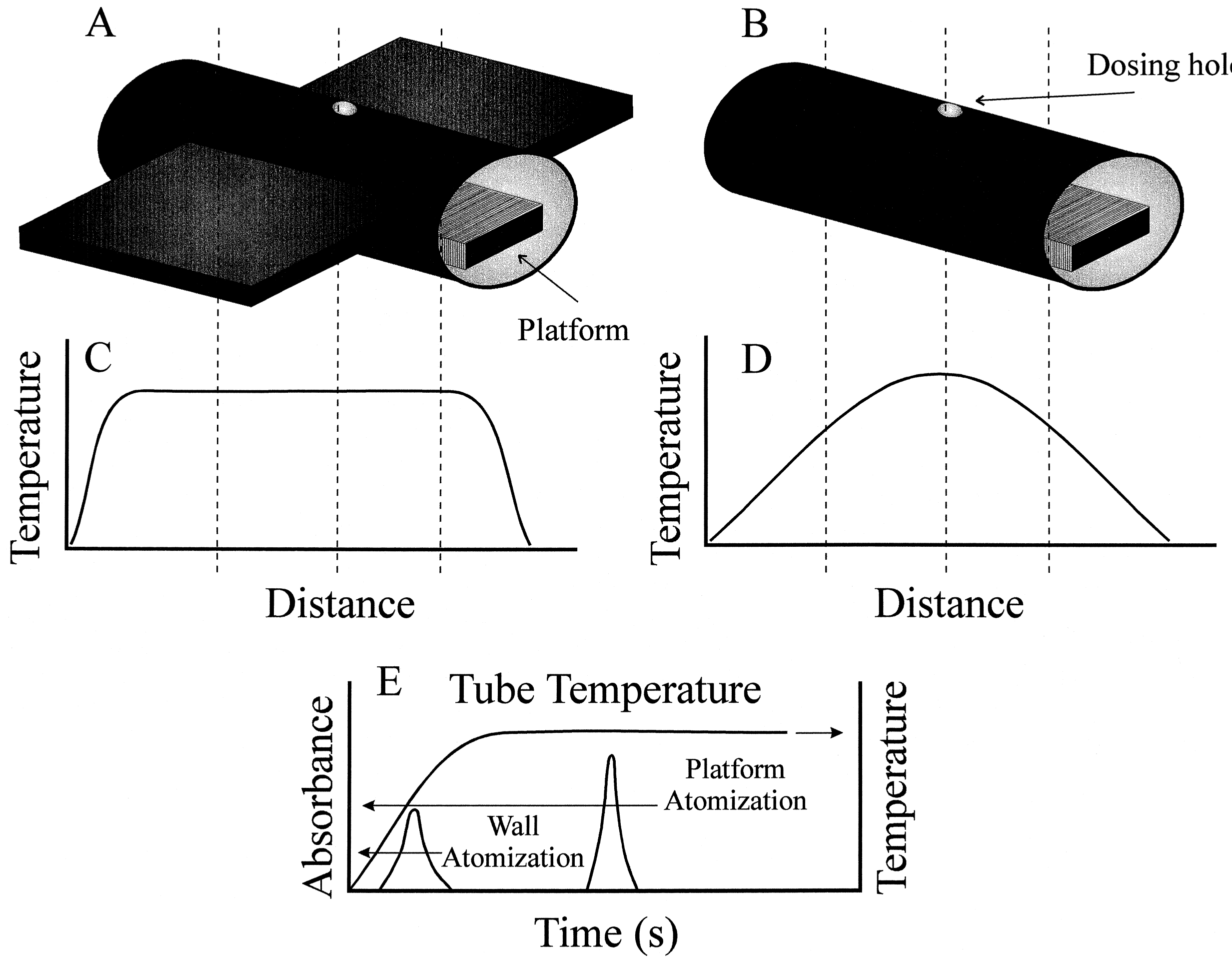


FIGURE 13 Graphite furnace electrothermal atomizers and their corresponding temperature profiles and analytical signals. (A) and (C) correspond to furnace geometry and the tube temperature profile for transversely heated atomizers. (B) and (D) correspond to furnace geometry and the tube temperature profile for longitudinally heated atomizers. The appearance time of the analytical signal for sample atomization from the furnace wall or the platform is highlighted in (E).

sample is pulse vaporized and the atomic vapor diffuses rapidly out of the tube.

2. Tube Design and Heating Characteristics

Figure 13 illustrates two popular tube designs for electrothermal vaporization. The transversely heated graphite tube (Fig. 13A), also known as an integrated contact cuvette, is heated through the electrical contacts at the side of the atomizer tube. This allows for a more uniform heating profile along the length of the graphite tube, as shown in Fig. 13C. In longitudinally heated furnaces (Fig. 13B), the atomizer is heated through the electrical contacts at the ends of the graphite tube. Although simpler in design, these furnaces tend to have a significant thermal gradient along the length of the graphite tube (Fig. 13D). Consequently, because the analyte is vaporized at the center of the tube, it may condense at the cooler ends of the graphite furnace.

Regardless of the furnace design, the sample is usually vaporized from a small graphite platform that is placed inside the graphite tube (Fig. 13A and B). The sample is placed on the platform through a small hole in the top center of the furnace tube. Figure 13E illustrates the different atomization profiles expected when the sample is vaporized either from the furnace wall or the platform. This figure helps to illustrate the function of the platform and explain why it is a useful tool. The graphite tube takes a finite time to heat up, as indicated by the tube-wall temperature versus time plot. The vapor temperature in the furnace follows the tube-wall temperature, and if the sample is placed on the bottom wall of the tube, it heats up at the same rate as the tube. In most cases, the sample is vaporized while the tube is still heating up and results in the

appearance of the analyte transient quite early in the heating cycle. The main problem in the early appearance of the atomization profile is with the gas-phase chemistry. The sample is vaporized into a cooler environment, which may not allow for decomposition of small molecules yielding a sufficient number of free atoms. But, if the sample is deposited on a small graphite platform (inside the graphite tube), it will be vaporized into the hot environment of the furnace later than in the wall atomization case. The platform heats up later than the tube because of its thermal mass and little thermal contact with the tube. As a result, the vaporized sample enters a much hotter environment than in the wall atomization case. The platform often provides a sufficient delay in vaporization so that many samples, upon vaporization, see a constant vapor temperature even though each matrix may release the analyte at slightly different times. One analytical scenario in which the platform vaporization may not be suitable is when refractory elements are to be determined. Refractory elements have a very high vaporization temperature, and platform atomization will result in premature furnace failure due to excessively high temperatures.

In a typical atomization cycle, the sample is vaporized with a heating rate of about 1300–1500 K/s. After reaching the desired temperature, the furnace is held at a constant temperature for 3–4 seconds while the absorption pulse is measured.

3. Pyrolytic Coating of the Graphite Tube

Atomic absorption spectrometry furnace tubes are supplied with a hard pyrolytic graphite surface. This helps to reduce the porosity of the tube, which minimizes adsorption of hot metal vapors and slows down the rate of oxidation of the graphite and the accompanying mechanical degradation of the furnace tube.

4. Inert Gas Requirements

An inert gas is used to flush the furnace during analysis. This prevents oxidation of the furnace. Usually an argon flow of about 300 cm^3/min. is maintained through the furnace during the dry and pyrolysis stages of the analysis. The flow is stopped or reduced during the atomization stage to minimize the gas expulsion from the furnace, which would reduce the absorbance signal.

5. The Detection System

The transient signal is usually measured quantitatively by evaluating the peak area (integrating) rather than by a measurement of peak height. While interference may affect position and height of the absorption peak, the area of the transient often does not change for a given analyte concentration. Because of the fast heating rates employed in electrothermal vaporization, the transient from the furnace lasts for a very short time and it is necessary to have an instrument with relatively fast (ms time-scale) electronics.

6. Physico-Chemical Interferences

The processes that lead to breakdown of the matrix, vaporization, and complete release of the analyte as a free element vary from analyte to analyte and matrix. These differing conditions are ultimately responsible for the interferences experienced by the analyte. These effects are well studied and the role of ubiquitous species such as carbon and oxygen are well documented with regard to the vaporization process. Even though an attempt is made to maintain an oxygen-free environment, the matrix usually contains oxygen. In addition, oxygen ingress through the dosing hole has been observed even during the atomization stage. Compounds of the analyte, such as oxides or halides (formed due to the presence of the matrix) may not break down sufficiently to release all the analyte into the source radiation.

Interference effects can often be minimized by the use of chemicals, which are added to the sample to modify the matrix so that interferences are controlled. A matrix modifier can both modify the matrix so that the analyte is vaporized relatively late during the atomization stage and allow a higher pyrolysis temperature to be used without premature release of the analyte. Matrix compounds can then be broken down during the pyrolysis step so that they are less likely to cause interference during the atomization stage. Some typical matrix modifiers are ascorbic acid, magnesium nitrate, and palladium solutions. Modifiers are added in relatively high amounts (about 250 μg).

7. Background Correction

Small molecules, such as sodium chloride, that are present during the atomization stage of a furnace analysis are often able to absorb or scatter radiation from the light source and cause a spurious absorption signal. This is called *spectral band interference*. In addition, smoke and other particles that are present during atomization can cause scatter of the radiation. This again causes an apparent spurious absorption signals. These signals can be corrected effectively by the use of one of two different methods of background correction.

Deuterium lamp background correction (two-source background correction): Two light sources are used in a deuterium lamp background corrected instrument. The first is the usual hollow cathode lamp, and the second is the deuterium lamp. The latter emits light over a broad range of wavelengths. The deuterium lamp does not normally

give a significant atomic absorption signal from the analyte element; however, it does give a scatter signal from particles within the furnace and an absorption signal from molecular species. The modulation chopper is mirrored so that, as the chopper spins, it alternatively irradiates the furnace by using either the HCL or the deuterium lamp. Hence, when the HCL is irradiating the furnace, a signal is obtained that is composed of the true analyte absorption signal plus the spurious background signal. When the deuterium lamp is irradiating the furnace, only the spurious background signal is obtained. The subtraction of these two signals gives the actual analyte absorption signal. The deuterium lamp works best for elements that have analytical wavelengths in the ultraviolet region of the spectrum. For background correction in the visible region a tungsten–halide lamp may be used.

Zeeman background correction: This method of background correction makes use of the fact that a magnetic field is able to split the atomic energy levels of an analyte. A magnet is placed around the furnace, with the lines of force of the field perpendicular to the direction of propagation of the light beam. In this configuration the atomic energy levels of an analyte can be split from the normal situation into three energy levels (σ_1, σ_2, and π). The π energy level absorbs only light of a particular polarization, which is a natural consequence of the splitting of the energy levels. The light source is deliberately polarized so that no light can be absorbed by the π energy level. Then two measurements are made. First, the magnetic field is switched off, and a measurement of signal plus background (A) is obtained. In this case, the light-source polarization is irrelevant because atoms that are not in a magnetic field absorb light of any polarization. Second, the magnetic field is switched on, causing the atoms to split into the various components. When the field is on, no analyte absorption of light occurs, because the light source is not of the correct depolarization even though it clearly is still at the correct wavelength. The σ components are not involved because they are at the wrong wavelength to absorb energy from the light source. Therefore, the signal obtained (B) results only from light absorbed or scattered by background species and particles. The simple subtraction (A–B) then gives the background-free analyte signal. This method has been proven to give reliable and accurate background correction for all elements that are normally analyzed by AAS. It is particularly easy to apply because only one light source is used which removes many alignment problems that causes difficulties in a two-source system. The background measurement is made at exactly the same wavelength as the source and analyte wavelengths. In contrast, two-source correction measures the background over the range of wavelengths that exit the monochromator.

V. ATOMIC FLUORESCENCE SPECTROMETRY

Atomic fluorescence is an extremely sensitive technique for determination of elements in samples. We should reiterate that in atomic fluorescence an external light source is used to excite the analyte atoms. An ideal light source for AFS must be much more intense than a hollow cathode lamp to achieve improvements in sensitivity. As a result, pulsed hollow cathode lamps and lasers are frequently used in AFS measurements. Excitation with a light source such as a hollow cathode lamp, which only emits radiation specific for the element of interest, makes AFS virtually completely free from spectral interferences. In addition, AFS is like AES in that a multi-element analysis can be achieved by putting several light sources around the atom cell, as discussed below.

A. Theoretical Background

This discussion assumes that the spectral line width of the light source is narrow relative to the absorption profile of the analyte atoms, as illustrated in Fig. 12A. The atoms absorb light from the source, and some of the energy is re-emitted as fluorescence, while various collisional processes in the atom cell deplete the remainder of energy. The ratio of the amount of light absorbed to that emitted is called the *quantum efficiency* (Y) and is ideally equal to one. In its simplest form, the equation for fluorescence radiant power (Φ_f) resembles general expressions in atomic absorption (because for optically unsaturated systems the fluorescence signal is a function of the initial source radiant power, Φ_0):

$$\Phi_f = \Phi_0(1 - e^{-kl}) \qquad (7)$$

In the above equation, k is the absorption coefficient and l is the optical path length for the atom cell. Therefore, in AFS the signal size is directly proportional to both the light-source intensity and the atom concentration. Calibration curves for AFS with HCL excitation are linear with a slope of 1 (on a logarithmic plot) at low concentrations and bend back towards the concentration axis with a limiting slope of –0.5 at high concentrations. The curvature at high concentration is related to self-absorption in the atom cell.

1. Light Source

The multi-element capability of AFS is realized by use of several light sources. The HCLs used for AFS are special high-intensity versions of the ones used for atomic absorption and up to 12 of them can be arranged in a practical experimental arrangement. Lasers are also frequently used

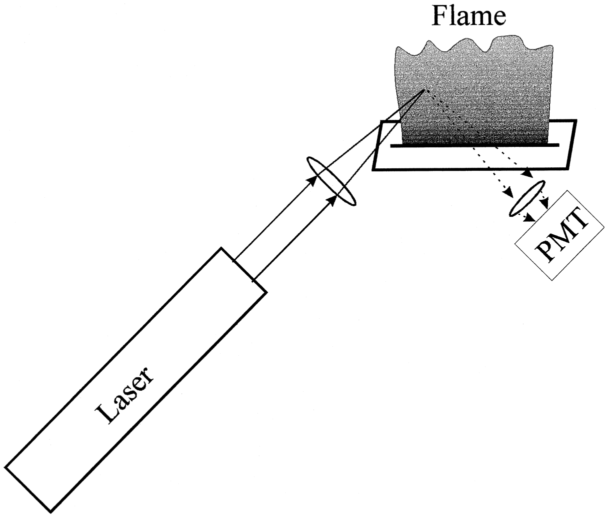


FIGURE 14 Laser atomic fluorescence in flame atomizers. The photomultiplier tube (PMT) is equipped with a band-pass filter.

for AFS measurements because of their inherently large radiant intensity. The general diagram for a flame atomic fluorescence instrument with laser excitation is shown in Fig. 14.

2. Detection System

Generally, monochromators are not used in AFS measurements. For hollow cathode excitation, each HCL has a paired dedicated photomultiplier tube detector. In front of each photomultiplier tube is a filter, which allows a range of wavelengths to pass through it including the atomic-fluorescence wavelength of the element excited by the HCL. The filter provides discrimination against the background emission from the atom cell. A high-resolution monochromator is not used for AFS, because the resolution is provided by the specificity of the light source for the element of interest. Although a low-resolution monochromator can be, and often is, used for AFS, filters pass more total light onto the photomultiplier tube and provide adequate resolution (in the range 2–10 nm) to discriminate against much of the background from the plasma. The light sources used for AFS measurement are always modulated to allow discrimination against background emission.

3. Spectral Interferences

Because of the selectivity of the excitation source, the spectral interferences are almost completely absent. Nonetheless, some spectral interference has been reported. This interference is caused by scatter of the incident source radiation off large droplets in the flame and, without background correction, could be mistaken for fluorescence. When an ICP is used as the atom cell, the scattering interference is not observed. The ICP is so efficient at breaking down droplets and particles that very few scatter signals have been detected in an HCL–ICP instrument.

VI. DETECTION OF ELEMENTAL IONS

The laser-enhanced ionization (LEI) technique or elemental mass spectrometry may be used to directly measure elemental ions. In LEI, a laser beam is used to excite the analyte atoms in either a flame or electrothermal atomizer cell. These atom cells are selected for LEI experiments because of their inherently low background ion population (in contrast to plasma sources, which are rich in ions). The laser-excited atoms undergo collisional processes that will ultimately lead to ionization of the analyte atoms. The ions are then collected and measured against a background current. Although extremely sensitive, the LEI technique is limited to detection of one analyte at a time in a flame or ETV atom cell. There are currently no commercial instruments available for LEI analysis.

Another approach for detection of elemental ions is to use ICP sources (which are very rich in ion population) in tandem with mass spectrometric detection. Under typical operating conditions, about half of the elements in the periodic table are singly ionized with an efficiency of 90% or greater. The schematic diagram for an ICP–MS instrument is shown in Fig. 15.

Unlike the upright configuration in ICP–AES, in ICP–MS the plasma torch is placed on its side along the axis of the mass spectrometric ion collection optics. A water-cooled cone is placed into the tip of the plasma to sample a portion of the ions generated by the ICP. The extracted ions pass through another orifice, the skimmer, through a differentially pumped region of the mass spectrometer. Differential vacuum pumping allows for the transfer of ions from plasma at atmospheric pressure into the high vacuum environment of the mass analyzer. The ions are guided through a series of lenses for the optimal signal-to-noise ratio at all mass ranges.

TABLE III Examples of Isobaric Interferences Observed in ICP–MS

Analyte	Mass	Interfering ion
Si	28	$^{12}C^{16}O^+$, $^{14}N_2^+$
K	39	$^{38}ArH^+$
Ca	40	$^{40}Ar^+$
As	75	$^{40}Ar^{35}Cl^+$
Se	80	$^{40}Ar_2^+$

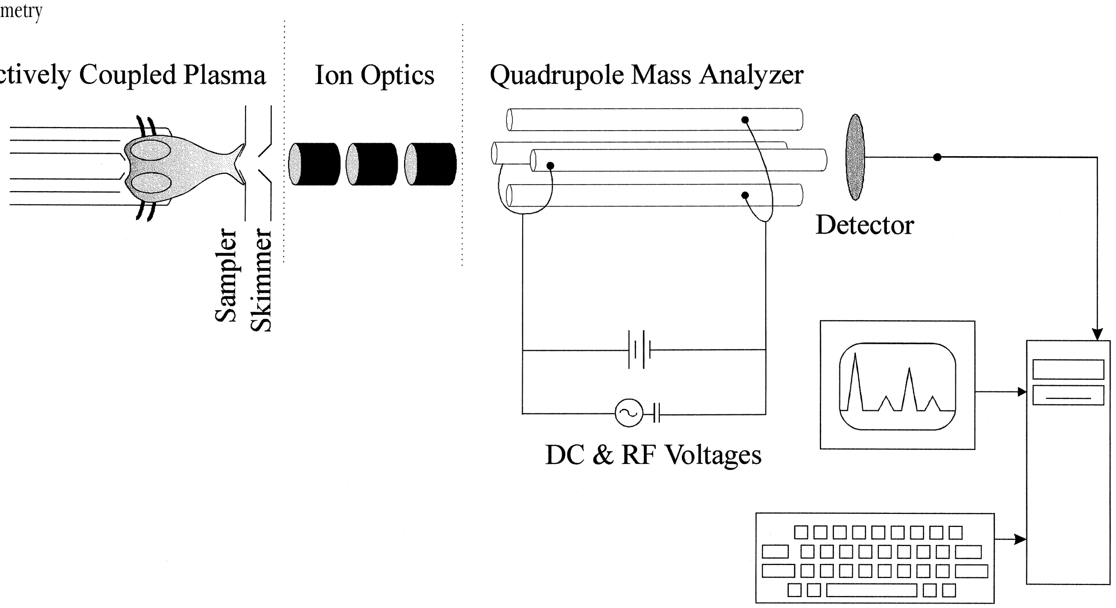


FIGURE 15 Schematic representation of an inductively coupled plasma mass spectrometer.

The linear dynamic range of an ICP–MS instrument is typically about 4–6 orders of magnitude, while the sensitivity is often 3 orders of magnitude better than the ICP–AES technique. While the detection limits for many elements with ICP–MS are better or comparable to AFS and graphite furnace atomic absorption spectrometry, the simultaneous multi-elemental detection afforded by ICP–MS clearly places this technique in the lead for analytical utility and versatility.

Interferences in ICP–MS are similar to chemical interferences observed for ICP–OES. In addition, a number of isobaric interferences caused by molecules with similar mass-to-charge ratios of some analyte ions also influence the analytical results. These interferences can be minimized by judicious use of acids in sample matrices, optimized plasma power settings, use of different sample introduction schemes, and use of altermate mass analyzers. Some common isobaric interferences observed in ICP–MS are listed in Table III.

VII. COMPARISONS OF THE ATOMIC SPECTROMETRIC TECHNIQUES

It is unlikely that any one of the above analytical techniques is capable of replacing all other methodologies. Some techniques, however, have all but disappeared (e.g., flame atomic emission), and others have been employed for unique projects (i.e., LEI, for selective isotopic detection). Frequently, a technique is selected based on the specific analytical needs and instrument availability. A general comparison of these techniques is outlined in Table IV. The information provided in this table is a gross oversimplification of many experimental factors and it should be

TABLE IV Comparison of Atomic Spectroscopic Techniques

	ICP–OES	dc arc	dc plasma	GD plasma	FAA	GF–AA	AFS	ICP–MS
Number of elements measured	E	E	E	E	A	A	E	E
Multi-elemental capability	E	E	E	E	P	P	A	E
Solids analysis	P	E	P	E	P	A	P	P
Micro samples	P	A	P	P	P	E	P	A
Qualitative analysis	E	E	E	E	P	P	P	E
Isotopic information	P	P	P	P	P	P	P	E
Detection limit	A	P	A	A	A	E	E	E
Dynamic range	E	P	A	A	P	P	A	E
Spectral (isobaric) interference	A	P	A	E	E	E	E	A
Chemical interference	E	P	A	A	A	A	A	E

Note: *E = excellent, A = acceptable, P = poor.

used only as a first approximation to narrow down the selection criteria.

SEE ALSO THE FOLLOWING ARTICLES

Analytical Chemistry • Atomic Physics • Mass Spectrometry

BIBLIOGRAPHY

Holcombe, J. A., Hieftje, G. M., and Majidi, V. (1998). "Focus on Analytical Spectrometry: A Compendium of Applied Spectroscopy Focal Point Articles," Society for Applied Spectroscopy, MD.
Ingle, J. D., and Crouch, S. R. (1988). "Spectrochemical Analysis," Prentice Hall, Englewood Cliffs, NJ.
Laitinen, H. A., and Ewing, G. W. (1981). "A History of Analytical Chemistry," American Chemical Society Maple Press, PA.

Auger Electron Spectroscopy

C. L. Briant

Division of Engineering, Brown University, Providence, RI

I. Introduction
II. The Auger Process and Auger Electron Spectroscopy
III. Applications in Materials Science
IV. Conclusions

GLOSSARY

Auger transition Electronic transition that produces Auger electrons. It occurs in the following way: and electron beam impinges on the surface of a solid and knocks out a core electron; an outer shell electron decays into the hole and in so doing emits energy; this energy can then knock another electron, the Auger electron, out of its orbital, and it can escape from the solid into the vacuum if the atom is located near the surface.

Grain boundaries Internal interfaces in solids where two single crystal grains meet one another. In most solids, the atomic matching between the interfaces is poor.

Intermetallic compound A compound made up of two or more metals. The atoms of each metal occupy specific sites on the crystal lattice. The compound is thus said to be ordered.

Segregation The process by which the composition of a surface or internal interface becomes enriched in a particular atomic species. The element reaches the interface by diffusion through the bulk.

AUGER ELECTRON SPECTROSCOPY is an electron spectroscopy that employs Auger, or secondary, electrons to determine the chemical composition of a solid surface. Because the electrons that are detected are of very low energy, this spectroscopy probes only the top two to five atom layers of the surface. This type of spectroscopy is now routinely used in materials research to obtain information about the composition of solid surfaces.

I. INTRODUCTION

Of all the techniques that have been developed to analyze surfaces, Auger electron spectroscopy has had the most widespread application. In the field of materials science, it has joined such analytical methods as X-ray diffraction and transmission electron microscopy as a staple of any well-equipped laboratory. It is used in chemistry and materials science to study the composition of solid surfaces and the chemical states of atoms and molecules on those surfaces. Chemists and physicists study the basic Auger transition to help learn about electronic processes in solids. Those interested in developing electronic equipment have been concerned with providing spectrometers with ever-decreasing incident beam diameters that will allow the chemical analysis of a surface on a microscopic scale. It is hoped that this article plus the

bibliography will provide the interested reader with a start in this field.

An obvious question that must be answered is why Auger electron spectroscopy has achieved this place in the surface analytical field. First of all, it is a relatively straightforward technique to use, and high-quality commercial spectrometers are available. The data are easily interpretable, at least on a qualitative level, and it is also possible to use an incident beam that is well under 1 μm in diameter. This diameter is several orders of magnitude smaller than what is possible for most other surface analytical techniques, and this fact means that Auger spectroscopy can be used to probe changes in composition across a surface at a microscopic scale. Finally, Auger electron spectroscopy can detect every element except hydrogen and helium, so it useful for studying both light and heavy element. We now wish to consider Auger electron spectroscopy and the process on which it is based.

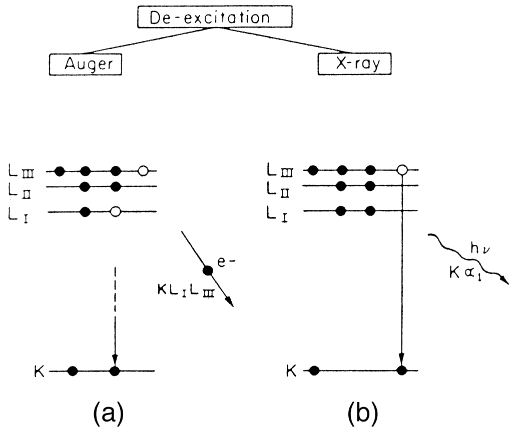


FIGURE 1 Energy level diagram representing de-excitations by (a) Auger electron emissions and (b) X-ray fluorescence. In the particular Auger process that is shown, a K-shell hole is first created and an electron in the L_I shell drops down to fill it. In so doing, it gives off enough energy to knock an electron out of the L_{III} shell, which becomes the Auger electron.

II. THE AUGER PROCESS AND AUGER ELECTRON SPECTROSCOPY

When an electron beam impinges on a solid surface, it can knock out inner shell electrons if it is of sufficient energy. The holes that are left behind will then be filled by having outer shell electrons fall into them. This last process occurs with the emission of energy, since the outer shell electron is changing from a higher to a lower energy. The energy that is given off can cause one of two events to occur. The energy can simply escape from the solid as an X-ray; in this case, the process of X-ray fluorescence has occurred. In the other case, this energy can knock out another outer shell electron. These electrons are the Auger electrons that were named for Pierre Auger who discovered this process in 1929. Such processes are usually denoted by the shells of the electrons that are involved. Thus, a KLL Auger process means that the first electron that was knocked out was from the K shell, the hole was filled by an electron from an L shell, and the Auger or secondary electron that was emitted was also from an L shell.

Figure 1 shows the processes of Auger emission and fluorescence schematically. The frequency with which one of them will occur relative to the other depends on the particular element; but as a rule of thumb, it can be remembered that elements with low atomic numbers tend to favor the Auger process, whereas those with high atomic numbers favor fluorescence. The Auger electrons that are emitted have energies that are characteristic of the element from which they came. Therefore, if they can be detected they will provide a method of chemical analysis. Furthermore, since the electrons that arise from such a process are going to be of low energy, those that do escape from the solid must come from very near the surface of the solid. Although the yield of Auger electrons will again depend on the particular element, it is estimated that most come from the top two to five atom layers.

It had been recognized for some years after their discovery that Auger electrons might provide a chemical analysis of a surface, but it was not until the work of L. A. Harris in 1968 that the technique actually became feasible in the laboratory. Prior to Harris' work, the problem had always been that the Auger signal was so small compared with that from other types of emitted electrons that it could not be adequately detected. Harris applied a phase-sensitive detection scheme in which a small, regularly oscillating voltage was superimposed on a larger, constantly increasing voltage. By measuring the electrons being collected, he was able to detect a small perturbation in the number of electrons at a given energy level, because the increased yield would be in phase with the frequency of the oscillating voltage. By electronically differentiating the signal in this way, he could help bring out these small peaks. With this advance, the use of Auger electrons to measure surface composition became a reality. Today, one can simply obtain differentiated spectra such as those shown in Fig. 2 and identify the elements present on the surface by comparing the energies of the peaks with those in handbooks of elemental standards.

Although Harris' contribution was pivotal in the development of Auger electron spectroscopy, there have been other advances since his that have helped Auger electron spectroscopy to become such a useful surface analytical technique. One of the most important was the development of the cylindrical mirror analyzer by Palmberg, Bohn, and

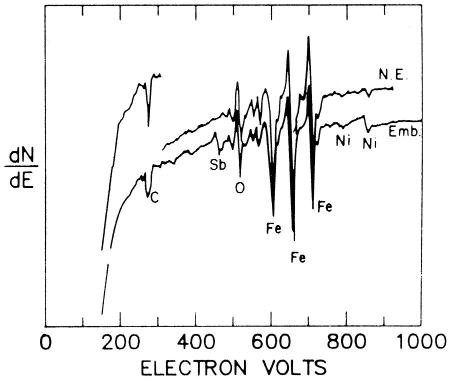


FIGURE 2 Typical differentiated Auger spectra taken from steels. The upper spectrum was from a nonembrittled (N.E.) steel, and the fracture surface that was analyzed was transgranular. The lower spectrum was from an embrittled steel, and the fracture surface that was analyzed was intergranular. [From Stein, D. F., Joshi, A., and LaForce, R. P. (1969). *ASM Trans. Q.* **62,** 776.]

Tracy. This analyzer provided a much higher signal-to-noise ratio and also allowed for much faster acquisition of data. It is the analyzer on which most modern systems are based. Another important improvement was in vacuum systems. When the early work was performed, it was almost impossible to obtain a clean surface. The vacuums were poor enough so that carbon and oxygen were adsorbed rapidly onto the surface. For a technique that only probes the top two to five atom layers, as does Auger electron spectroscopy, this adsorption presented a problem. With the development of ions pumps and turbomolecular pumps that provide a background pressure of less than 1×10^{-10} Torr, a clean surface could be prepared routinely, either through fracture or by sputtering with rare gas ions, and kept in this condition for 5 to 10 h. Finally, new and much brighter filaments were developed that allowed a stronger Auger signal to be generated. In particular, the development of the lanthanum hexaboride filament, which replaced the standard thermionic tungsten filament, offered a great increase in filament current. This increase also meant that smaller areas (<1 μm in diameter) on the surface could be analyzed, and it led to the development of the scanning Auger spectrometer. In this spectrometer, the Auger beam is first used to form a secondary electron image of the surface that is being analyzed. The beam can then be placed on one particular area to obtain a spectrum from that spot. In the most recent machines, field emission filaments are used that allow analysis of a spot size of under 20 A.

One problem with Auger electron spectroscopy is the difficulty of converting the signal to accurate values of atomic percent. The reason for this difficulty is that because the Auger signal is so surface sensitive, it is very dependent on such factors as trace contamination, the crystallography of the surface, the roughness of the surface, and the presence of other alloying elements in the solid that is being analyzed. Although one can obtain accurate calibration for a particular system, it is very difficult to apply a given calibration to a variety of alloys on a routine analytical basis. Consequently, many researchers frequently rely on relative changes in a normalized signal to discuss changes in the surface composition. For example, in the differentiated spectrum shown in Fig. 2, which was taken from a steel fracture surface, one might choose the strong iron peak at 703 eV and normalize all other peaks by it. Then, one could discuss changes in normalized peak height ratios to describe changes in the composition of this surface.

The previous paragraphs give a brief introduction of the Auger process and the basic way in which surfaces are analyzed with it. However, the great power of Auger electron spectroscopy has been broad in its applicability. We now wish to discuss these applications in the field of materials science.

III. APPLICATIONS IN MATERIALS SCIENCE

Auger electron spectroscopy has become one of the most important tools for the investigation of interfaces in the field of materials science. In fact, it is probably safe to say that it has found its most widespread application in this field. Because it does not give detailed chemical bonding information about the species on a surface, other techniques such as XPS and UV photoemission are more commonly used for detailed chemical investigations. However, in materials science applications the need is more often to detect the presence of elements on the surface of the solid, and for this application Auger electron spectroscopy is the preferred technique.

The types of studies that have been performed can be divided into two categories. One set is concerned with studies of the external surface. Segregation to the surface can cause its composition to be very different from that of the bulk. There has been great interest in understanding how this difference in composition can affect such processes as catalysis, corrosion, and oxidation. Also, one can use the kinetics of segregation to the surface to determine the diffusion coefficient of an element in the solid. The other type of study is concerned with internal interfaces, such as grain boundaries and particle–matrix interfaces. The composition of these interfaces can also be quite different from that of the bulk, as a result of segregation, and consequently, they can control various metallurgical properties.

Let us first consider the external surfaces. The most common types of experiments that are performed are those in which the surface is cleaned and then heated to allow segregation. The heating allows solid state diffusion to occur, and one finds that with time the composition of the surface can be significantly altered as a result of this segregation. Figure 3 shows an example of such changes in composition. The base material is an Fe-Cr-Ni 304 stainless steel that has bulk concentrations of phosphorus, sulfur, and nitrogen that are below 300 ppm. However, as seen in Fig. 3, the concentration of these elements on the surface can be quite significant.

The most common use of these types of studies has been to simply detect elements present on the surface during a surface science experiment or to determine what elements will tend to segregate to interfaces when the materials are heated. Some investigations have attempted to use surface segregation to predict the segregation that will occur to internal interfaces such as grain boundaries, but there are numerous studies which show that the differences between grain boundary segregation and surface segregation are great enough that such studies can, at best, give qualitative trends.

One of the concerns of this type of experiment is that it must be done in high vacuum, whereas most applications are in an environment of some type. Thus, one must be careful in the application of the results of these experiments to situations where aggressive media are the cause of the problem, such as would be the case for corrosion or high-temperature oxidation. Another word of caution must be given about the analysis of surfaces that have been exposed to the atmosphere. Because of the extreme surface sensitivity of Auger spectroscopy, there are two ways

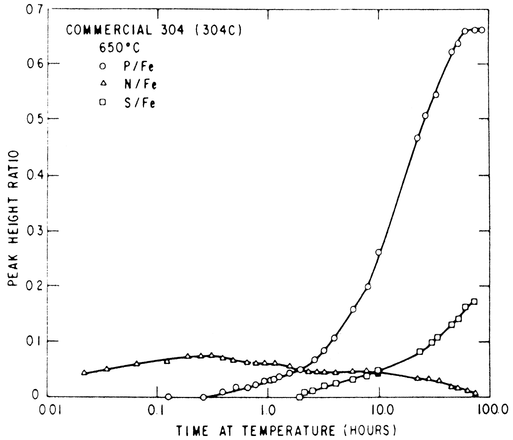


FIGURE 3 The segregation kinetics at 650°C for sulfur, phosphorus, and nitrogen in a commercial 304 stainless steel alloy. In this figure, changes in the peak height ratio are used to express changes in the composition of the surface. [From Briant, C. L., and Mulford, R. A. (1982). *Metall. Trans. A*, **13A,** 745.]

in which results could be misinterpreted. One is that the surface will be covered with enough oxygen and carbon from the environment to mask the surface species that may have originally caused a problem. The second is that contaminants from handling of the material may show up and mislead the researcher in his or her effort to find the root cause of a problem.

In contrast, the direct studies of segregation to internal interfaces have been extremely important in the solution of many engineering problems. The basic segregation process is similar to that which occurs for surface segregation. The solid is heated to a sufficiently high temperature and solid state diffusion begins to occur. However, there is an important difference in the way grain boundary segregation is detected compared with the way surface segregation is studied. For the surfaces, the change in the composition is simply monitored by impinging the electron beam on the surface while it is being heated. However, grain boundaries are internal interfaces and they must be exposed after the segregation has occurred. What is done is to take advantage of the fact that segregation often weakens the grain boundaries. The samples can then be fractured apart along their grain boundaries in the high-vacuum spectrometer, and this fracture surface becomes the one that is analyzed by Auger electron spectroscopy. However, the fracture may not be completely intergranular, and this application is one where the development of scanning Auger electron spectrometers has provided a great advantage. One can obtain an image of the fracture surface and select regions on it that are intergranular and analyze only them. One can then compare the compositions of these regions with those from regions where the fracture goes through the grains.

The reason why Auger analysis of grain boundaries has been so important is because the segregation often controls metallurgical properties. We have already mentioned that segregation often makes the boundaries preferred fracture paths, and, in general, intergranular fracture is a very low energy, undesirable fracture mode. Prior to the development of Auger electron spectroscopy, it was suspected that segregation of impurity elements, such as phosphorus, tin, and antimony, caused the embrittlement. Figure 4 shows the change in the ductile to brittle transition temperature of Ni-Cr steel plotted as a function of the amount of segregation of phosphorus and antimony. This transition temperature measures the brittleness in steels, and as the transition temperature increases, the brittleness of the steel increases. This plot shows that there is a good correlation between brittleness and segregation.

Another example of the effect of segregation can be to affect the response of the material to an external environment. Numerous studies have shown that phosphorus segregation to grain boundaries of steel enhances its corrosion nitrate environments, nitric acid, picric acid, and

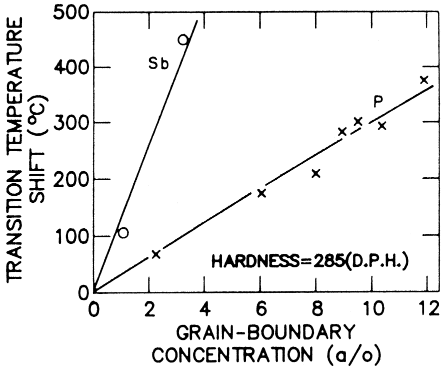


FIGURE 4 The change in the ductile to brittle transition temperature plotted as a function of the grain boundary concentration of either phosphorus or antimony. The steels were a Ni-Cr base material and were doped with one of these two elements. The grain boundary concentration was determined by Auger electron spectroscopy. [From Mulford, R. A., McMahon, C. J., Jr., Pope, D. P., and Feng, H. C. (1976). *Metall. Trans. A*, **7A,** 1269]

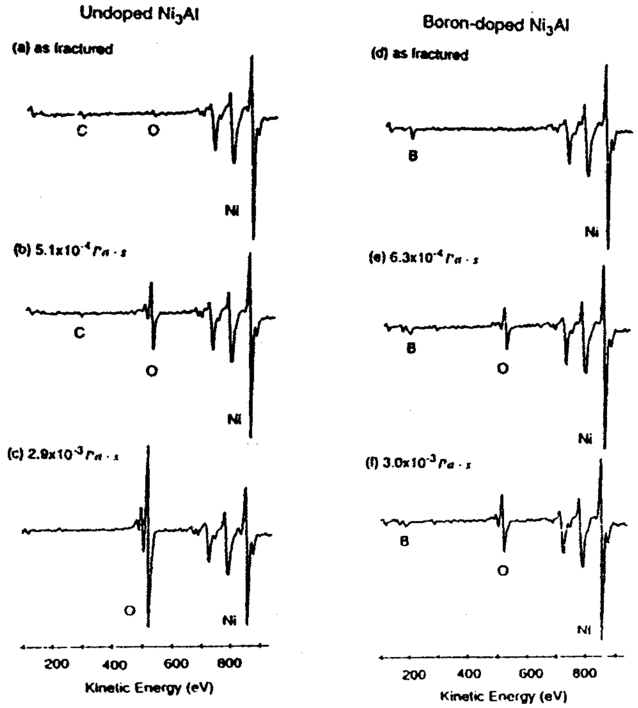


FIGURE 5 Auger spectrum from intergranular fracture surfaces of undoped (a), (b), and (c) and boron-doped (d), (e), and (f) Ni_3Al. (a) and (d) are spectra from as-fractured surfaces. All other samples were exposed to the indicated pressure of water vapor. [From Lee, K. H., Lukowski, J. T., and White, C. L., (1996). *Scripta Metall.* **35,** 1153.]

caustic solutions. Furthermore, it has been well established that segregants that embrittle grain boundaries also make the material more susceptible to hydrogen embrittlement.

Another type of environmental embrittlement that has received significant attention in recent years has been the embrittlement of intermetallic compounds when they are exposed to water vapor. In general, it has been found that the water-molecules decompose on the surface of these compounds and release atomic hydrogen. This hydrogen then diffuses into the solid and causes embrittlement. Research has also shown that the ease with which the dissociation of water vapor occurs depends on the chemistry of the surface.

As an example of this effect, consider the results reported in Fig. 5. The intermetallic compound under investigation was Ni_3Al, and the particular issue was to determine whether or not the presence of boron on the grain boundaries affected the dissociation of water vapor. The interest in boron additions stems from the fact that boron on the grain boundaries of this compound appears to enhance grain boundary cohesion. For this study, the measure of dissociation of the water vapor on the surface was the height of the oxygen Auger peak. (Recall that hydrogen does not have an Auger signal.) To perform the experiment the sample was fractured and exposed to water vapor. Figure 5 shows two sets of results. Those on the left are for a sample that did not contain boron. Those on the right are for a sample that did contain boron. The results clearly show that the presence of boron on the exposed grain boundaries decreased oxygen absorption and presumably dissociation of water vapor. Thus, this study suggests that one reason why boron enhances cohesion at the grain boundaries is because it retards hydrogen embrittlement.

These examples are just a few of the many applications that Auger electron spectroscopy has had in materials science. The references at the end of the article give many more of them. Yet it seems fair to say that the overall theme behind most of these studies is the fact that surfaces and internal interfaces can have very different compositions from that of the bulk and that this difference in composition causes them to have different properties. Auger electron spectroscopy is used to detect the composition of the interface and allow correlation between composition and properties to be made.

IV. CONCLUSIONS

This article has considered Auger electron spectroscopy. We have described the basic Auger transition, fundamental ways in which Auger electrons are detected, and various problems that can be encountered in using this spectroscopy. We have also considered the application of this spectroscopy in materials science. In the early days of the development of this technique, it was a specialized technique that was often the center of a particular study.

Today, primarily as a result of the commercial availability of excellent systems, it is used along with other techniques to solve complex engineering problems.

SEE ALSO THE FOLLOWING ARTICLES

Embrittlement, Engineering Alloys • Infrared Spectroscopy • Photoacoustic Spectroscopy • Raman Spectroscopy • Surface Chemistry • X-Ray Photoelectron Spectroscopy

BIBLIOGRAPHY

Briant, C. L., ed. (1999). "Impurities in Engineering Materials," Dekker, New York.

Briant, C. L., and Messmer, R. P., eds. (1988). "Auger Electron Spectroscopy," Academic Press, Boston.

Briggs, D., and Seah, M. P. (1983). "Practical Surface Analysis by Auger and X-ray Photoelectron Spectroscopy," Wiley, New York.

Smith, G. C. (1994). "Surface Analysis by Electron Spectroscopy: Measurement and Interpretation," Plenum, New York.

Thompson, M., Baker, M. D., Christie, A., and Tyson, J. F. (1985). "Auger Electron Spectroscopy," Wiley, New York.

Capillary Zone Electrophoresis

Tim Wehr
LC Resources, Inc.

I. Principles of Capillary Electrophoresis
II. Band Broadening in Capillary Electrophoresis
III. Capillary Electrophoresis Instrumentation
IV. Capillary Electrophoresis Separation Modes
V. Capillary Zone Electrophoresis
VI. Capillary Isoelectric Focusing
VII. Capillary Sieving Techniques
VIII. Isotachophoresis
IX. Micellar Electrokinetic Chromatography
X. Capillary Electrokinetic Chromatography

GLOSSARY

Anolyte Acidic solution (usually dilute phosphoric acid) used to terminate the anodic end of the pH gradient in isoelectric focusing.

Carrier ampholytes Synthetic amphoteric compounds containing acidic and basic functions which migrate under the influence of an electric field to form a stable pH gradient in isoelectric focusing. The composition of the carrier ampholyte mixture defines the extent and shape of the pH gradient.

Catholyte Basic solution (usually dilute sodium hydroxide) used to terminate the cathodic end of the pH gradient in isoelectric focusing.

Chiral selector Additives used in capillary electrophoresis and electrokinetic chromatography which complex differentially with enantiomers such that the complexes migrate at different rates.

Electroosmotic flow Bulk flow of the electrolyte when an electric field is applied to a capillary carrying fixed charges on the inner surface.

Isoelectric point The pH at which an amphoteric species (such as a polypeptide) has zero net charge and does not migrate in an electric field.

Joule heat The heating of a conductive medium as current flows through it.

Plate height In chromatography and electrophoresis, the measure of the amount of dispersion or spreading of a zone or peak during the separation process.

CAPILLARY ELECTROPHORESIS (CE) is a relatively new separation technology which combines aspects of both gel electrophoresis and high-performance liquid chromatography (HPLC). Like gel electrophoresis, the separation depends upon differential migration in an electrical field. Since its first description in the late 1960s, capillary electrophoretic techniques analogous to most

conventional electrophoretic techniques have been demonstrated: zone electrophoresis, isotachophoresis, isoelectric focusing, and sieving separations. Unlike conventional electrophoresis, however, the separations are usually performed in free solution without the requirement for a casting a gel. As in HPLC, detection is accomplished as the separation progresses, with resolved zones producing an electronic signal as they migrate past the monitor point of a concentration-sensitive (e.g., ultraviolet absorbance or fluorescence) detector; therefore, the need for staining and destaining is eliminated. Data presentation and interpretation is also similar to HPLC; the output (peaks on a baseline) can be displayed as an electropherogram and integrated to produce quantitative information in the form of peak area or height. In most commercial CE instruments, a single sample is injected at the inlet of the capillary and multiple samples are analyzed in serial fashion. This contrasts to conventional electrophoresis in which multiple samples are frequently run in parallel as lanes on the same gel. This limitation of CE in sample throughput is mitigated by the ability to process samples automatically using an autosampler. Several multicapillary systems have been introduced to increase sample throughput; most of these are designed specifically for high-throughput DNA sequencing.

Compared to its elder cousins, CE is characterized by high resolving power, sometimes higher than electrophoresis or HPLC. The use of narrow-bore capillaries with excellent heat-dissipation properties enables the use of very high field strengths (sometimes in excess of 1000 volts/centimeter), which decreases analysis time and minimizes band diffusion. When separations are performed in the presence of electroosmotic flow (EOF), the plug-flow characteristics of EOF also contribute to high efficiency. In contrast, the laminar flow properties of liquid chromatography increase resistance to mass transfer, reducing separation efficiency.

Because of its many advantages, CE is finding increasing use as a tool in analytical chemistry. In some cases it may replace HPLC and electrophoresis, but more often it is used in conjunction with existing techniques, providing a different separation selectivity, improved quantitation, or automated analysis. It was the adaptation of CE technology in the design of 96-channel DNA sequencers that enabled a draft of the human genome sequence to be completed in less than two years.

I. PRINCIPLES OF CAPILLARY ELECTROPHORESIS

As the name implies, capillary electrophoresis separates species within the lumen of a small-bore capillary filled with an electrolyte. A schematic of a CE system is presented in Fig. 1. The capillary ends are immersed in electrolyte-filled reservoirs containing electrodes connected to a high-voltage power supply. A sample is introduced at one end of the capillary (the inlet) and analytes are separated as they migrate through the capillary towards the outlet end. As separated components migrate through a section at the far end of the capillary, they are sensed by a detector, and an electronic signal is sent to a recording device.

As in conventional gel electrophoresis, the basis of separations in capillary electrophoresis is differential migration of analytes in an applied electric field. The electrophoretic migration velocity v_{EP} (in units of cm/sec) will depend upon the magnitude of the electric field, E, and electrophoretic mobility, μ_{EP}, of the analyte:

$$v_{EP} = \mu_{EP} E \quad (1)$$

In a medium of a given pH, mobility of an analyte is given by the expression:

$$\mu_{EP} = q/6\pi\eta r \quad (2)$$

where q is the net charge of the analyte, η is the viscosity of the medium, and r is the Stoke's radius of the analyte. Since the Stoke's radius is related to molecular mass and therefore to frictional drag, mobility will increase inversely with molecular weight and directly with increasing charge. Mobility in units of $cm^2\ sec^{-1}\ V^{-1}$ can be determined from the migration time and field strength by:

$$\mu_{EP} = (I/t)(L/V) \quad (3)$$

where I is the distance from the inlet to the detection point (termed the *effective length* of the capillary), t is the time required for the analyte to reach the detection point (migration time), L is the total length of the capillary, and V is the applied voltage.

In contrast to most forms of gel electrophoresis, the velocity of an analyte in capillary electrophoresis will also depend upon the rate of electroosmotic flow. This phenomenon is observed when an electric field is applied to a solution contained in a capillary with fixed charges on the capillary wall. Typically, charged sites are created by ionization of silanol (SiOH) groups on the inner surface of the fused silica. Silanols are weakly acidic and ionize at pH values above about 3. Hydrated cations in solution associate with ionized SiO^- groups to form an electrical double layer: a static inner Stern layer close to the surface and a mobile outer layer known as the Helmholtz plane. Upon application of the field, hydrated cations in the outer layer move toward the cathode, creating a net flow of the bulk liquid in the capillary in the same direction (Fig. 2). The rate of movement is dependent upon the field strength

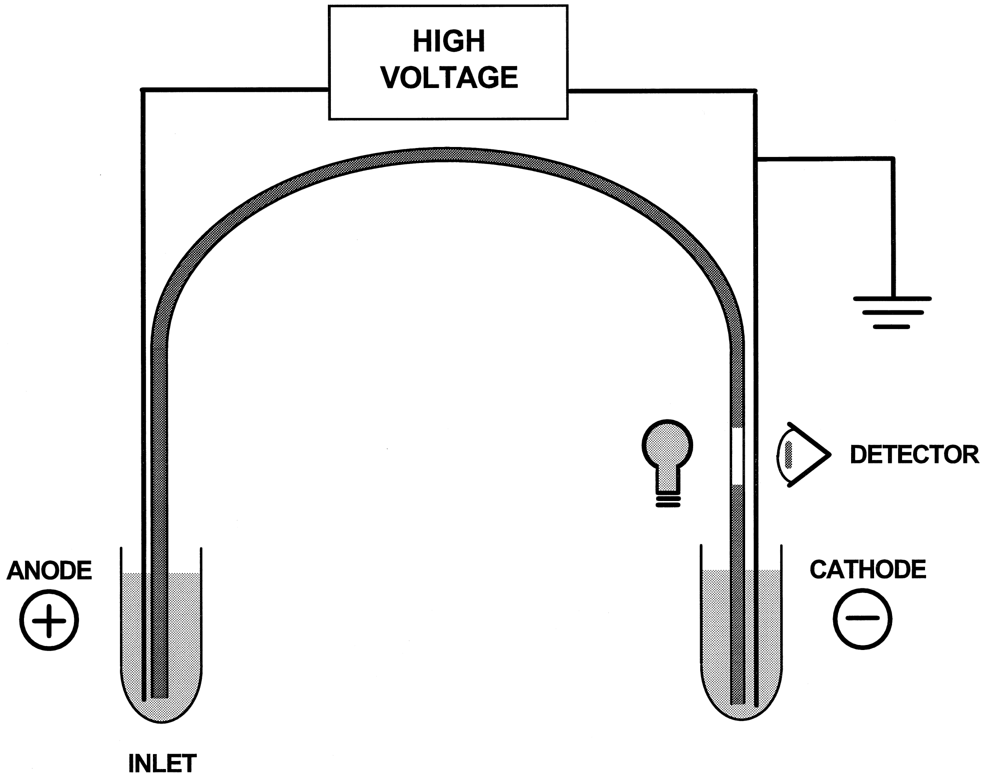


FIGURE 1 General schematic of a capillary electrophoresis system. Capillary internal diameters can range from 25 to 150 μm, and capillary lengths can range from 20 cm to over 1 meter. In some cases, sample may be injected at the capillary outlet, and the short segment between the outlet and the detection window serves as the separation path.

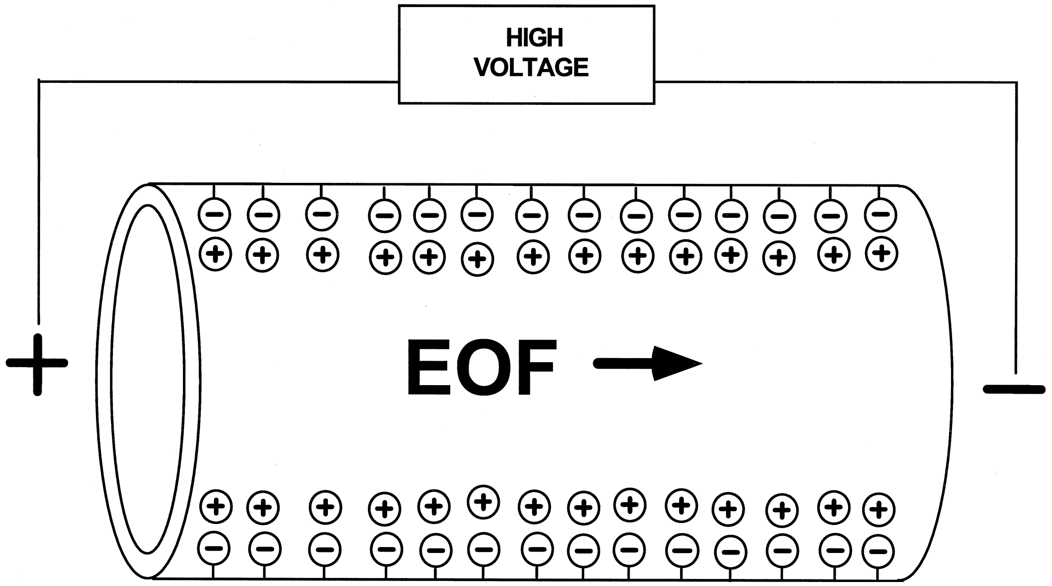


FIGURE 2 Electroosmotic flow in an uncoated fused silica capillary. The negative charges represent ionized silanol groups on the silica surface, and the positive charges represent hydrated electrolyte cations.

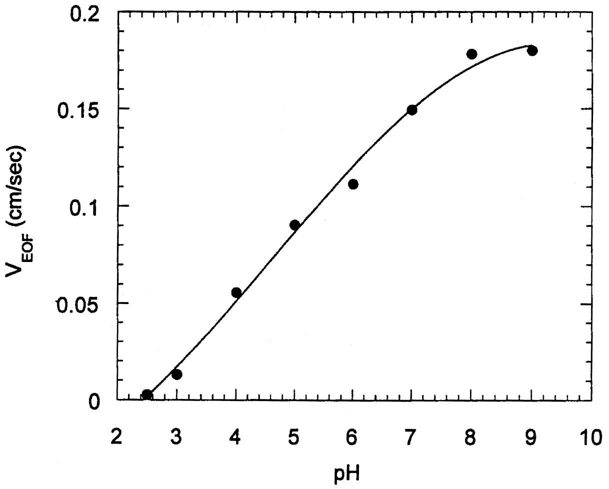


FIGURE 3 Magnitude of electroosmotic flow as a function of pH.

and charge density of the capillary wall. The population of charged silanols is a function of the pH of the medium, so the magnitude of EOF increases directly with pH until all available silanols are fully ionized (Fig. 3). The velocity ν_{EOF}, of electroosmotic flow can be expressed as:

$$\nu_{EOF} = \mu_{EOF} E \tag{4}$$

where μ_{EOF} is the electroosmotic mobility, defined as:

$$\mu_{EOF} = \varepsilon \xi_w / \eta \tag{5}$$

where ε is the dielectric constant, ξ_w is the zeta potential of the capillary wall, and η is the viscosity of the medium. Electroosmotic mobility can be determined experimentally by injecting a neutral species and measuring the time, t_{EOF}, when it appears at the detection point:

$$\mu_{EOF} = (\mathrm{I}/\mathrm{t}_{EOF})(\mathrm{L}/\mathrm{V}) \tag{6}$$

The apparent velocity, ν_{APP}, of an analyte in an electric field will therefore be the combination of its electrophoretic velocity and its movement in response to EOF:

$$\nu_{APP} = \nu_{EP} + \nu_{EOF} = (\mu_{EP} + \mu_{EOF})E \tag{7}$$

In the presence of electroosmotic flow, analytes which possess net positive charge will migrate faster than the rate of EOF, analytes which have zero net charge will be carried toward the cathode at the rate of EOF, and anionic analytes will migrate towards the cathode at a rate which is the difference between their electrophoretic velocity and ν_{EOF}. If the magnitude of EOF is sufficiently great, all analytes regardless of their charge state will migrate past the detection point. In this regard, performing separations in the presence of EOF is highly desirable. However, achieving reproducible separations requires that EOF be constant, and this in turn requires that the surface characteristics of the capillary wall remain constant from run to run. Compounds such as polypeptides and organic amines are notorious for interacting with silica surfaces and changing the level of EOF. A great variety of capillary surface treatments and buffer additives have been developed for reducing adsorption and controlling EOF, and this continues to be an active area of research in capillary electrophoresis.

Separation efficiency N in capillary electrophoresis is given by the following expression:

$$N = \mu_{APP} V / 2D \tag{8}$$

where D is the diffusion coeffcient of the analyte in the separation medium. This predicts that efficiency is only diffusion limited and increases directly with field strength. Note that $V = E/L$, so increasing capillary length will increase the diffusion time, decreasing efficiency. Capillary electrophoresis separations are usually characterized by very high effiency, often as high as several hundred thousand plates. Efficiency in CE is much higher than in HPLC because there is no requirement for mass transfer between phases and because the flow profile in EOF-driven systems is flat (approximating plug flow) in contrast to the laminar-flow profiles characteristic of pressure-driven flow in chromatography columns. Resolution, R, in CE is defined as:

$$R = 1/4\left(\Delta\mu_{EP} N^{1/2} / \mu_{EP} + \mu_{EOF}\right) \tag{9}$$

This implies that resolution will be greatest when μ_{EP} and μ_{EOF} are of similar magnitude but of opposite sign; however, high resolution will be at the expense of analysis time. It is also evident from Eq. (9) that, although Eq. (8) suggests that use of higher field strengths is the most direct route to high efficiency, a doubling of voltage yields only a 1.4-fold increase in resolution, at the expense of Joule heat. The most direct route to optimize resolution is to increase $\Delta\mu$.

II. BAND BROADENING IN CAPILLARY ELECTROPHORESIS

Band broadening and the resultant reduction in resolution can arise from several contributing factors. If the total band broadening is expressed as plate height H, the contributions to band broadening due to initial zone width, diffusion and electrodispersion, Joule heating, and adsorption can be expressed qualitatively as:

$$H = h_{inj} + h_{diff+cond} + h_{joule} + h_{ads} \tag{10}$$

A. Initial Zone Width (h_{inj})

Best resolution will always be obtained by keeping the initial sample zone as small as possible. The starting zone

length should not exceed 5% of the total capillary length. Sample zone sharpening can be achieved by preparing the sample in an electrolyte of lower conductivity than the analysis buffer. Under these conditions, there is a discontinuity in field strength at the sample:buffer boundary such that ions migrating rapidly from a region of higher field strength become focused at the boundary. This focusing or stacking effect not only produces narrow zones for increased resolution, but also increases zone concentration for enhanced sensitivity. Sensitivity enhancement is directly proportional to the ratio of sample to buffer conductivity. Sample zone sharpening can also be accomplished by transient isotachophoretic preconcentration, as discussed below.

B. Diffusion and Electrodispersion ($h_{diff+cond}$)

Band broadening due to axial diffusion should be reduced by shortening analysis time (i.e., by operating at high field strengths with short capillaries). On the other hand, higher field strengths will generate more Joule heat, which increases diffusion rates. Electrodispersive band broadening arises from conductivity differences between the zone and the background electrolyte. If this difference is large, diffusion at one boundary of the zone is neglible due to the zone-sharpening effect caused by the discontinuity in field strength. In this case, the peak will be asymmetric; that is, it will exhibit fronting or tailing depending on whether conductivity of the zone is greater or less than that of the background electrolyte. This phenomenon is readily observable when analyzing highly charged small molecules such as inorganic ions, and satisfactory peak shapes can only be obtained by carefully matching sample and electrolyte conductivities. The effect is much less noticeable in other CE separations, except at high sample concentrations.

C. Joule Heating (h_{joule})

When an electric field is applied to a capillary containing an electrolyte, Joule heat is generated uniformly across the circumference of the tube. Since heat can only be removed at the margin of the tube, a temperature gradient exists across the radius of the tube. As noted in Eq. (2), mobility is inversely related to viscosity, which decreases with temperature. Mobility increases approximately 2.5% for each degree rise in temperature. Therefore, the temperature gradient creates a mobility gradient across the tube radius which contributes to band broadening. It should also be noted that, in addition to this mobility gradient, Joule heat can compromise resolution by increasing diffusion and convection; these are minimized by the use of viscous buffers.

The small internal diameters of fused silica capillaries have been the key to high-resolution CE separations. The smaller the capillary bore, the greater the surface-to-volume ratio and the more efficiently heat is removed from the tube. The optimal capillary inside diameter (i.d.) for most applications is 50 to 75 μm. Capillaries with smaller diameters are subject to plugging, and the high surface-to-volume ratio increases the risk of adsorption. Capillaries with larger internal diameters may exhibit significant loss in resolution due to thermal effects, and may require operation at low field strengths or with low-conductivity buffers.

D. Adsorption (h_{ads})

Wall interactions have been the greatest obstacle to achieving satisfactory resolution and reproducibility in capillary electrophoresis, particularly in zone electrophoresis and isoelectric focusing. Approaches to minimizing or eliminating adsorption include use of buffer additives and dynamic or covalently-coupled capillary coatings.

III. CAPILLARY ELECTROPHORESIS INSTRUMENTATION

A variety of commercial CE instruments are available, from simple modular systems consisting of a power supply, detector, and injection device to fully integrated automated systems under computer control. This discussion will focus on the features of automated CE instrumentation.

A. Power Supply

Power supplies capable of delivering constant voltage with high precision up to 30 kV are standard throughout the industry. Most systems offer, in addition, constant current operation at up to 300 μA; constant current operation may be desirable in systems without adequate temperature control. High voltage is applied at the capillary inlet, with the outlet (detector) at ground potential. When separations are performed in uncoated fused silica capillaries in the presence of EOF, the inlet (high-voltage) electrode is the anode and EOF carries analytes towards the cathode (“normal” polarity). Separations performed in coated capillaries without EOF may require “reversed” polarity. In some separations, the direction of EOF may be reversed by using capillaries coated with positively charged polymers or by using osmotic flow-modifying additives in the background electrolyte. In these cases, reversed polarity is also used.

B. Injection

Sample injection in CE requires the introduction of very small amounts of analyte at the capillary inlet with high precision. All commercial instruments offer electromigration injection and at least one type of displacement injection.

Electromigration is the simplest injection method in CE; the capillary inlet is immersed in the sample solution and high voltage is applied for a brief period (typically a few seconds). If no electroosmotic flow is present, sample ions enter the capillary by electophoretic mobility alone; this mode of sample introduction is termed *electrophoretic injection*. If EOF is present, sample ions will be introduced by a combination of electrophoretic mobility and electroosmotic flow; this mode is generally termed *electrokinetic injection*.

Electrophoretic injection offers two advantages. First, if electrophoretic injection is performed in the absence of EOF, only species of like charge will enter the capillary. This enables discrimination against compounds of opposite charge, simplifying the separation problem. Second, zone sharpening can be achieved using the stacking principle described above. Unfortunately, these advantages are countered by two major limitations. Since sample ions enter the capillary based on mobility, low-mobility ions will be loaded to a lesser extent than high-mobility ions. More importantly, presence of non-analyte ions in the sample will reduce injection efficiency, so electrophoretic injection is very sensitive to the presence of salts or buffers in the sample matrix. The disadvantages of electrophoretic injection argue against its use in routine analysis except in cases where displacement injection is not possible (e.g., in capillary gel electrophoresis, CGE). Electrokinetic injection suffers from the further disadvantage that many sample matrices contain components such as proteins which adsorb to the capillary wall and change the magnitude of EOF.

Displacement or hydrodynamic injection is usually the preferred method since analyte ions are present in the sample zone in proportion to their concentration in the bulk sample, and injection efficiency is less sensitive to variations in sample ionic strength. However, it should be noted that the presence of high salt can affect detector response with displacement injection, and variations in the sample viscosity (due to temperature variations or the presence of viscosity-modifying components) can affect displacement injection efficiency.

Two modes of displacement injection have been employed in commercial CE instruments: application of positive pressure at the capillary inlet and application of vacuum at the capillary outlet. The former method can employ pressurization of the sample headspace by gas (pressure injection) or application of hydrostatic pressure by elevating the capillary inlet relative to the capillary outlet (gravity injection). Gravity injection has been reported to provide reproducible injection of very small sample zones. On the other hand, pressure injection can allow the flexibility of introducing larger sample zones; this can be an advantage when using injection pneumatics to introduce a chiral selector or to perform on-column concentration. Moderate injection pressures can be an advantage when injecting samples into capillaries containing analysis buffers with viscous agents such as sieving polymers.

C. Capillary Temperature Control

Temperature control of the capillary environment is essential for attaining satisfactory reproducibility. Inadequate temperature control results in variable migration times. In CE, peak area depends upon the residence time of the component in the detector light path and therefore is dependent upon migration velocity. If migration times vary because of inadequate temperature control, peak area precision will be poor. Control of capillary temperature above ambient temperature may be desirable in special applications—for example, in performing kinetic studies or on-column enzyme assays, in the study of protein folding, or for mutation detection in amplified gene sequences. Operation at higher temperatures in capillary zone electrophoresis will decrease analysis time and may improve peak shape. Operation at lower temperatures generally has little advantage in most modes of capillary electrophoresis.

Capillary temperature control can be achieved by forced air or nitrogen convection or by a circulating liquid coolant. Forced air control permits the use of free-hanging capillaries but is less efficient. Liquid cooling requires that the capillary be enclosed in a sealed cartridge, but the cartridge format provides for automatic alignment of the capillary in the detector light path and reduces time required for capillary installation when changing methods.

D. Detectors

Most of the detection modes used in HPLC have been demonstrated for capillary electrophoresis (absorbance, fluorescence, conductivity, electrochemical, radioactivity, mass spectrometry, postcolumn reaction). However, of these, only absorbance, fluorescence, conductivity, and mass spectrometry are available for automated CE systems.

1. Absorbance

As in HPLC, absorbance detection is used in the vast majority of CE applications, and all commercial CE systems

employ UV or UV–Vis absorbance as the primary mode of detection. The simplest approach is the use of line-source lamps or continuum-source lamps with wavelength selection by filters. Better flexibility is obtained using a continuum (e.g., deuterium lamp) source with wavelength selection by a grating monochromator.

All commercial CE absorbance detectors employ on-tube detection, in which a section of the capillary itself is used as the detection cell. This permits detection of separated zones with no loss in resolution. Most capillaries used for CE are coated with with a polymer (usually polyimide) which protects the fused silica capillary and provides it with mechanical stability. Since the polymer is not optically transparent, it must be removed from the detection segment to form a "window," and this window must be accurately positioned in the optical path to achieve good sensitivity. This segment of bare capillary is very fragile and is subject to breakage during manipulation and installation of the capillary. Capillaries with a UV-transparent coating are commercially available which eliminate this problem; however, the coating is not resistant to some coolants (e.g., fluorinated hydrocarbons) used in liquid-cooled CE systems.

In on-tube detection, the internal diameter of the capillary forms the detection light path. In accordance with Beer's law, the sensitivity of a concentration-sensitive detector is a direct function of the length of the light path. Therefore, in comparison to an HPLC detector with a 1-cm path length, detector signal strength should be reduced 200-fold in a CE system equipped with a 50-μm-i.d. capillary. Concentration sensitivity can be improved by employing focusing lenses to collect light at the capillary lumen, by detecting at low wavelengths (where most analytes have greater absorbance), and by using sample-focusing techniques during the injection process. However, even under ideal conditions, the concentration limit of detection (CLOD) is about $10^{-6}\,M$. Sensitivity in optical detectors can be improved by the use of extended-pathlength capillaries. These include capillaries with increased diameters in the window segment ("bubble cell" capillaries) or in-line cells with a section of capillary positioned axially along the light path ("Z-cells").

Several commercial CE systems incorporate scanning absorbance detectors. Scanning detection enables on-the-fly acquisition of spectra as analytes migrate through the detection point; this information can assist in the identification of peaks based on spectral patterns, in detection of peak impurities by variation in spectral profiles across a peak, or in determation of the absorbance maximum of an unknown compound. Two different designs are used to accomplish scanning detection in CE instruments. In photodiode array (PDA) detectors, the capillary is illuminated with full-spectrum source light; the light passing through the capillary is dispersed by a grating onto an array of photodiodes, each of which samples a narrow spectral range. In fast-scanning detectors, monochromatic light is collected from the source using a movable grating and slit assembly and directed to the capillary; light transmitted by the capillary is detected by a single photodiode. Scanning is accomplished by rapidly rotating the grating through an angle to "slew" across the desired spectral range. Photodiode array detectors are more sensitive in full-scan mode, but the high light intensity can damage coated or gel-filled capillaries.

2. Fluorescence

Fluorescence detection offers the possibility of high sensitivity and, in the case of complex samples, improved selectivity. However, this mode of detection requires that the analyte exhibit native fluorescence or contain a group to which a fluorophore can be attached by chemical derivatization. The number of compounds that fall into the former category are small, and while many analytes contain derivatizable groups (e.g., amino, carboxyl, hydroxyl), most derivatization chemistries are limited by one or more disadvantages (slow reaction kinetics, complicated reaction or cleanup conditions, poor yields, interference by matrix components, derivative instability, interference by reaction side products or unreacted derivatizing agent).

When compared to fluorescence detectors for HPLC, the design of a fluorescence detector for CE presents some technical problems. In order to obtain acceptable sensitivity, it is necessary to focus sufficient excitation light on the capillary lumen. This is difficult to achieve with a conventional light source, but is easily accomplished using a laser. The most popular sources for laser-induced fluorescence (LIF) detection are the argon ion and helium–neon lasers, which are stable and relatively inexpensive. These lasers produce emission lines close to the excitation wavelengths for several common fluorophores. The CLOD for a laser-based fluorescence detector can be as low as $10^{-12}\,M$.

The most successful applications of CE–LIF detection have been in the analysis of carbohydrates and nucleic acids. Carbohydrates can be derivatized with agents such as aminopyrene trisulfonic acid, which introduces a fluorophore and also confers electrophoretic mobility to the analyte. Double-stranded nucleic acids can be detected using intercalating dyes and by covalent attachment of nucleotides carrying fluorophores. The fluorescently labeled primers or chain terminating nucleotides which are widely used in automated slab gel-based DNA sequencers are also used in capillary-based sequencers. In one multi-capillary sequencer, a novel sheath flow system is used for LIF excitation of separated bands with imaging of emitted fluorescence on a charge-coupled device (CCD) detector.

3. Mass Spectrometry

On-line coupling of a capillary electrophoresis system to a mass spectrometer (CE–MS) enables mass and structural information to be obtained for separated components. The most common configuration is introduction of the capillary outlet into an electrospray ionization (ESI) interface. In this configuration, the outlet electrode of the CE is eliminated and the MS becomes the ground. Two types of CE–ESI interfaces have been employed. For conventional capillaries of 50- to 100-μm i.d., a makeup flow is introduced via a coaxial sheath flow or liquid junction coupling to transport the capillary eluant into the ionization source and produce a stable electrospray. In these designs, the concentric stainless steel capillary serves as the electrical connection for the electrospray voltage. An alternative design, termed microspray or nanospray, permits direct introduction of the CE eluant without makeup flow. Small diameter capillaries are used, with the outlet tip tapered and metal-coated to establish the ESI high-voltage connection. The major limitation of CE–ESI/MS is the requirement for volatile buffers, which narrows the choice of CE separation modes and resolving power. However, this constraint may be relaxed to a degree by the recent introduction of orthogonal or "Z-spray" ESI designs. In some studies, an in-line, hollow-fiber dialysis tube has been used between the separation capillary and the ESI interface. This technique has been used to remove ampholytes in the performance of capillary isoelectric focusing–ESI–MS of proteins.

E. Preparative Capillary Electrophoresis

Because of its high resolving power, CE is often considered for micropreparative isolation. Many of the commercially available CE systems have the capability for automatic fraction collection; however, the desirability of using CE as a preparative tool has to be carefully weighed against the problems encountered in fraction collection. When using narrow-bore capillaries, the volume injected into the capillary is quite small (typically a few nanoliters). Unless the analyte is in very high concentration, recovery of sufficient material requires repetitive injections of the same sample. In this case, the run-to-run migration times must be highly reproducible to ensure accurate collection of the analyte peak. Also, the recovered analyte must be stable under the collection conditions for the time required to collect the desired amount of material (often several hours). An alternative strategy is the use of larger diameter capillaries ($\geq 75\ \mu$m). However, thermal effects may compromise resolution, and low voltages or low-conductivity buffers may be necessary to prevent excessive heating.

IV. CAPILLARY ELECTROPHORESIS SEPARATION MODES

One of the major advantages of CE as a separation technique is the wide variety of separation modes available. Analytes can be separated on the basis of charge, molecular size or shape, isoelectric point, or hydrophobicity. The same CE instrument can be used for zone electrophoresis, isoelectric focusing, sieving separations, isotachophoresis, and chromatographic techniques such as micellar electrokinetic chromatography and capillary electrokinetic chromatography. The following sections provide a brief description of each separation mode and typical applications.

V. CAPILLARY ZONE ELECTROPHORESIS

In capillary zone electrophoresis (CZE), the capillary, inlet reservoir, and outlet reservoir are filled with the same electrolyte solution. This solution is variously termed "background electrolyte," "analysis buffer," or "run buffer." In CZE, the sample is injected as a plug at the inlet end of the capillary, and sample components are resolved into separated zones according to their mass-to-charge ratio as they migrate towards the detection point. It is the simplest form of CE and the most widely used.

A. Separation Conditions for Capillary Zone Electrophoresis

Capillary zone electrophoresis may be performed in the presence or absence of EOF. Separations carried out in the presence of EOF permit the analysis of cations and anions in the same analysis, and the resolution of analytes moving counter to EOF can be enhanced. Uncoated fused silica capillaries exhibit electroosmotic flow toward the cathode, and the magnitude of EOF increases with pH. Capillaries coated with an adsorbed or covalent positively charged coating exhibit reversed electroosmotic flow toward the anode. Reversed-charge CZE can be used for analysis of cationic analytes which would adsorb to bare silica capillaries, and for analysis of high-mobility anions which would not be resolved with normal EOF flow. A disadvantage of performing CZE in the presence of EOF is variation in the magnitude of EOF, which can compromise migration time and peak area reproducibility. Electroosmotic flow can be controlled or eliminated by coating the capillary with an adsorbed or covalent neutral material, usually a hydrophilic polymer. Coated capillaries are often used when analyzing species that have strong affinities for silica, such as proteins.

Mobility and selectivity in CZE are most profoundly affected by analyte charge, and selection of the electrolyte pH is the most effective method of controlling a CZE separation. A wide variety of buffers have been employed in CZE, and a buffer is selected to provide good buffering capacity at the desired pH, low UV absorbance, and low conductivity. In addition to the buffer, other components may be added to the electrolyte to control EOF, reduce solute–wall interactions, or to modulate the mobility or solubility of an analyte. Additives for CZE include neutral salts, organic amines, surfactants, organic solvents, and chiral selectors. Secondary equilibria introduced by additive–analyte interactions are very important for achieving resolution in CZE.

B. Applications of Capillary Zone Electrophoresis

The great flexibility of CZE has enabled its successful application to a wide variety of analytes ranging from inorganic ions to macromolecules. A very incomplete listing of these includes analysis of ionic drugs and metabolites in pharmaceutical formulations and body fluids; analysis of foods, food supplements, and nutriceuticals; analysis of milk and cereal proteins; quality control and stability monitoring of protein therapeutics; clinical analysis of serum proteins; and quality control of wines. The CZE separation of peptides originating from tryptic digestion of cytochrome C is shown in Fig. 4. In addition to these applications, three important variations of CZE are of particular note.

1. Capillary Ion Analysis

Inorganic ions are a particular challenge for CZE because of their high mobilities and lack of UV absorbance. Inorganic ions can be detected in the UV region using indirect detection. In this technique, a UV-absorbing species is used as the background electrolyte, and analytes are detected as negative peaks. High-mobility anions are analyzed using reversed-charge CZE. Capillary ion analysis is routinely used in the analysis of environmental samples, industrial brines, and food products.

2. Affinity Capillary Electrophoresis

Affinity capillary electrophoresis (ACE) is used to characterize receptor–ligand interactions. The technique depends upon a shift in the electrophoretic mobility of the receptor upon complexation with a charged ligand. The

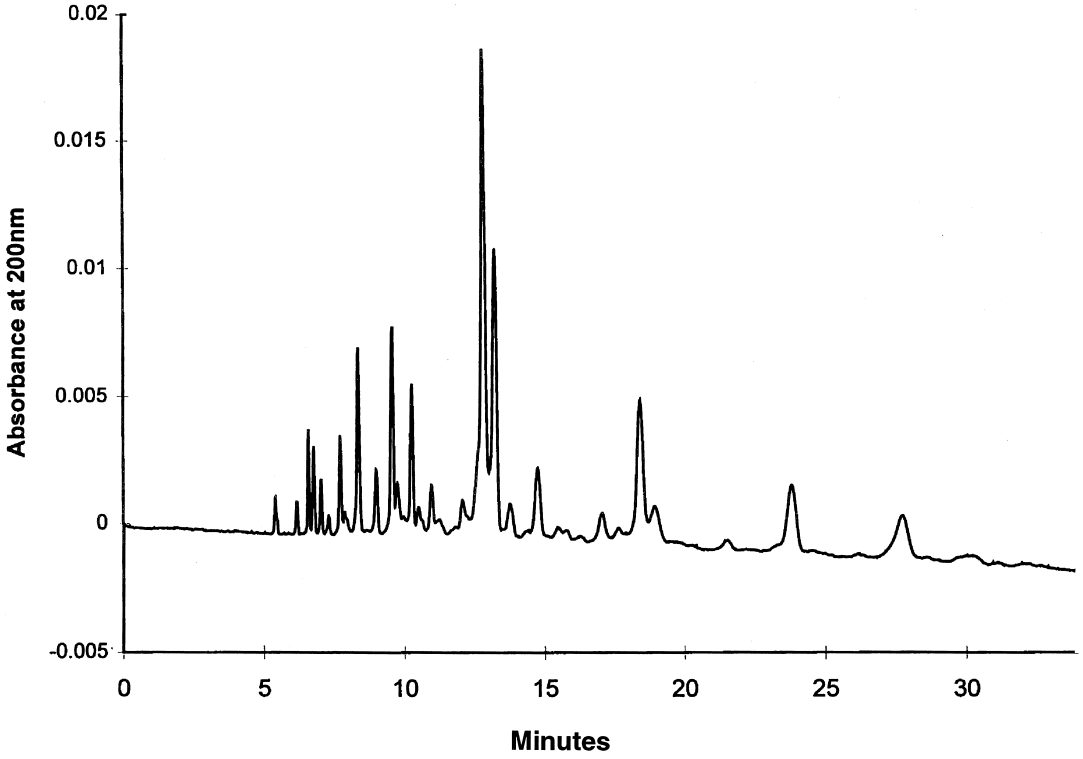


FIGURE 4 Separation of a tryptic digest of cytochrome C by capillary zone electrophoresis. Separation was performed in a neutral-coated capillary using 100-m*M* sodium phosphate (pH 2.5).

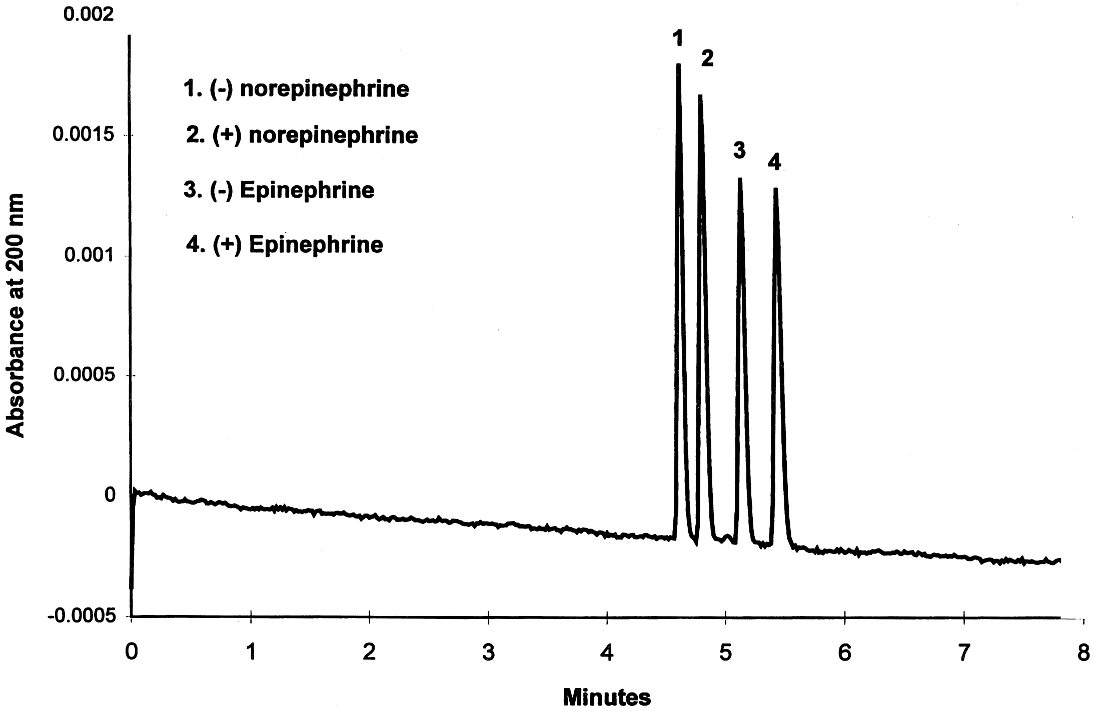


FIGURE 5 Separation of enantiomers of epinephrine and norepinephrine. Separation was performed in a neutral-coated capillary using 100-m*M* sodium phosphate (pH 2.5) and β-cyclodextrin.

distribution of bound vs. free receptor at various ligand concentrations can be used to determine binding kinetics. Applications of ACE include determination of enzyme–effector binding, antibody–antigen binding, and DNA–carcinogen interactions.

3. Chiral Analysis

Enantiomers can be resolved by incorporating a chiral selector in the background electrolyte. Typical chiral selectors are cyclodextrins, proteins such as α_1-acid glycoprotein, and macrocyclic antibiotics such as vancomycin. Charged enantiomers may be separated by differential formation of inclusion complexes with neutral cyclodextrins, while uncharged enantiomers may be resolved using a charged cyclodextrin such as cyclodextrin-sulfobuytlether. Separation of enantiomers of epinephrine and norepinephrine using a cyclodextrin chiral selector is shown in Fig. 5.

VI. CAPILLARY ISOELECTRIC FOCUSING

Isoelectric focusing (IEF) is an equilibrium technique in which amphoteric compounds are focused in a stable pH gradient at the point where they have no net charge. The pH gradient is contained between a basic catholyte solution and an acidic anolyte solution. The technique is almost universally applied to separation of proteins, which are resolved according to their isoelectric points. Capillary isoelectric focusing (CIEF) is similar in concept to conventional gel IEF; carrier ampholytes form a pH gradient under the influence of an electric field and proteins migrate electrophoretically to the point in the gradient at which they are isoelectric. In both IEF formats, very high resolution is achieved due to the electrophoretic focusing effect, which counters protein diffusion away from the zone at equilibrium. In contrast to gel IEF (in which focused proteins are fixed and detected by staining *in situ* within the gel), capillary IEF with on-tube detection requires mobilization of focused protein zones past the detection point.

A. Separation Conditions for Capillary Isoelectric Focusing

Capillary IEF can be carried out as a single- or two-step process. In single-step CIEF, a mixture of sample proteins and ampholytes is introduced as a plug at the capillary inlet, and an electric field and a mobilizing force are applied simultaneously. Formation of the pH gradient and protein focusing occur while the zone is in transit from the capillary inlet to the detection point. The mobilizing force can be pressure, vacuum, gravity, or electroosmotic. Although single-step CIEF is simple in concept, achieving

reproducible migration patterns can be difficult, particularly when using electroosmotic mobilization.

In two-step CIEF, focusing and mobilization are performed separately. Initially, the entire capillary is filled with a mixture of sample and ampholytes, high voltage is applied, and the focusing progress is monitored by the decrease in operating current. Once equilibrium has been established, the mobilization step is begun. The mobilizing force may be pressure, vacuum, gravity, or electrophoretic. In all cases, the electric field is maintained to prevent defocusing of zones during mobilization. Electrophoretic mobilization is accomplished by changing the composition of the anolyte or catholyte, usually by the addition of a salt, buffer, or zwitterion solution. This change in composition produces a progressive shift in pH down the capillary, causing electrophoretic migration of focused proteins past the detection point (Fig. 6). In two-step CIEF, the presence of EOF disrupts the CIEF process, resulting in loss of resolution and reproducibility. For this reason, two-step CIEF is performed in capillaries coated to reduce or eliminate EOF. Because the entire capillary is filled with sample in two-step CIEF, zone concentrations at equilibrium will be higher and detection sensitivity will be greater compared to single-step CIEF.

Resolution in CIEF is governed primarily by the composition of the ampholyte mixture. For screening and for analysis of complex mixtures of proteins, wide-range ampholyte blends are used which generate gradients of over 8 pH units (e.g., pH 3–10); for high-resolution separation of a few proteins, narrow-range ampholyte blends are used to generate gradients of 1 to 2 pH units.

B. Applications of Capillary Isoelectric Focusing

The high resolving power of CIEF enables separation of protein species with small differences in isoelectric points. Capillary IEF has been applied to the analysis of protein glycoforms, protein clips, and deamidation products. The technique has also been used for analysis of hemoglobin variants and glycated hemoglobins. It provides a rapid and convenient method for estimating the isoelectric point of an unknown protein. On-line CIEF–MS is being investigated as an instrumental counterpart to analytical two-dimensional gel electrophoresis in proteomic studies.

VII. CAPILLARY SIEVING TECHNIQUES

Sieving techniques are required for separation of species which have no differences in mass-to-charge ratio such as nucleic acids and sodium dodecylsulfate (SDS)–protein complexes. Sieving systems include cross-linked or linear polymeric gels cast in the capillary or replaceable polymer solutions.

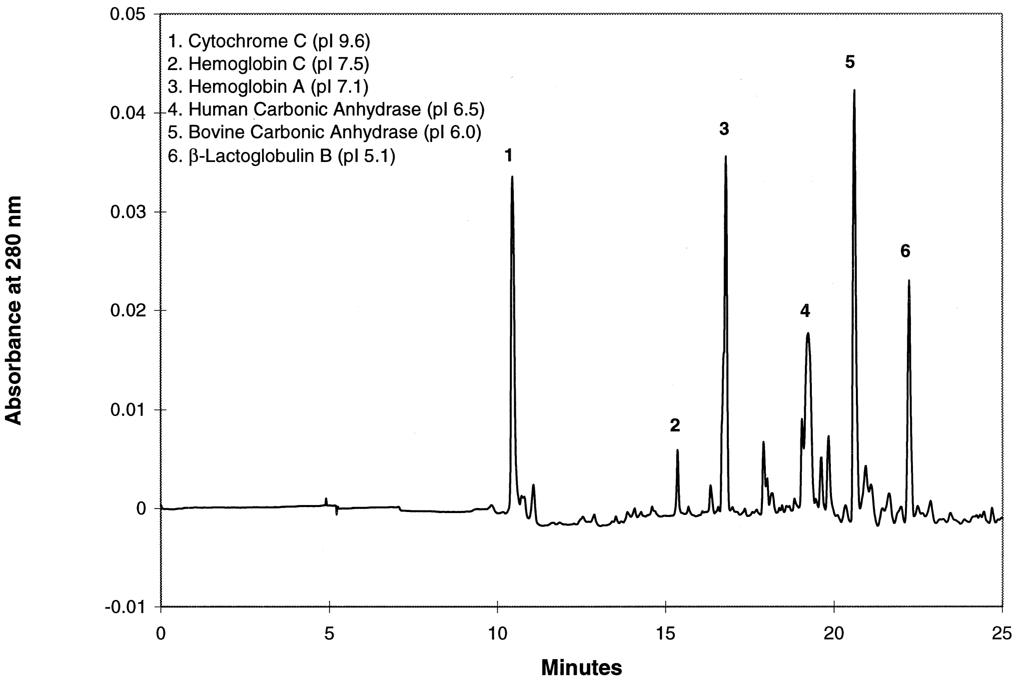


FIGURE 6 Separation of proteins by two-step CIEF using electrophoretic mobilization. The ampholytes generated a gradient from pH 3 to 10 after a focusing time of 300 sec in a neutral-coated capillary. Cathodic mobilization was initiated by replacing the catholyte (40-m*M* NaOH) with an alkaline zwitterion solution. pI, Isoelectric point.

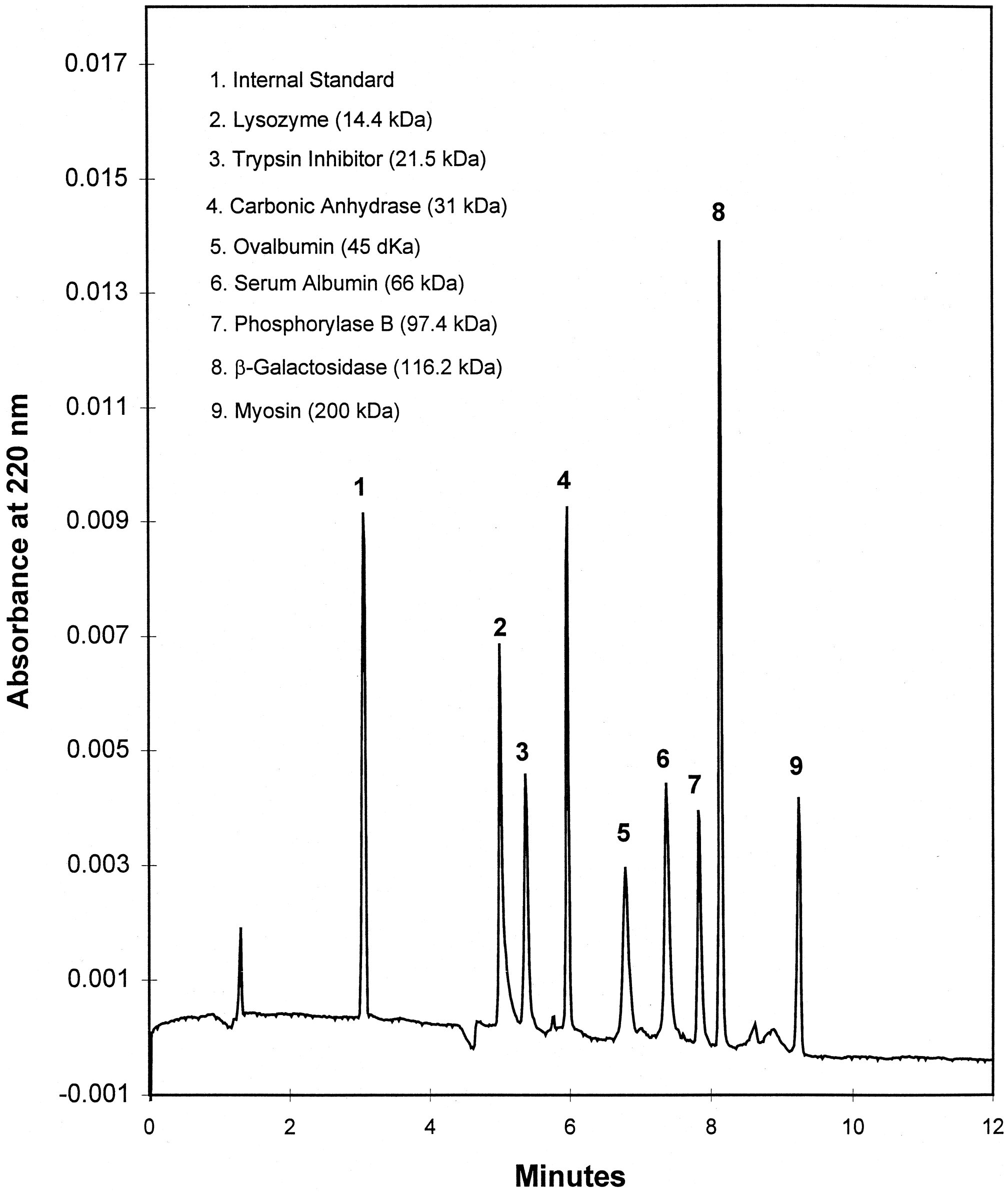


FIGURE 7 Separation of SDS–protein complexes by polymer-sieving CE. Separation was obtained in an uncoated capillary using a commercial polymer-sieving applications kit developed for analysis of SDS–protein complexes (Bio-Rad Laboratories; Hercules, CA).

A. Capillary Gel Electrophoresis

Polyacrylamide gels can be cast within a capillary using the same chemistry employed for casting conventional polyacrylamide gels. However, preparation and use of gel-filled capillaries is complicated by bubble formation caused by volume changes during polymerization, and by extrusion of the gel from the capillary. To circumvent this, gels are covalently attached to the capillary wall and polymerized in a segmental fashion. Care must be exercised in their use to avoid bubble formation and contamination of the gel bed. An advantage of gel-filled capillaries is the ability to closely control the porosity by changing the ratio of monomer to cross-linker. Polyacrylamide gel-filled capillaries are incompatible with absorbance detection at short UV wavelengths. Capillary gel electrophoresis (CGE) has been used for high resolution of synthetic antisense nucleotides such as phosphothioate nucleotides.

B. Replaceable Polymer Solutions

Sieving separations can be accomplished using solutions of linear or branched hydrophilic polymers. These

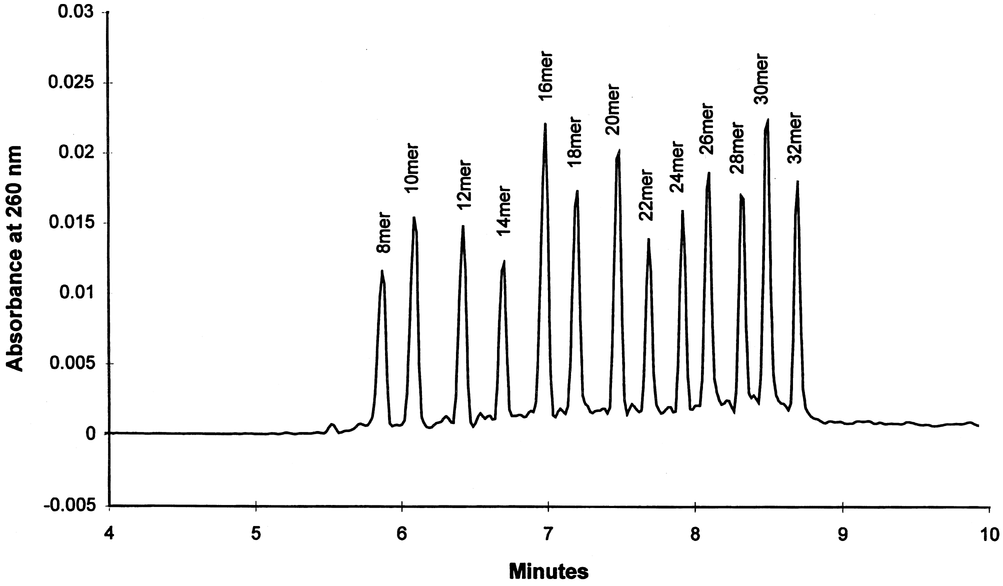


FIGURE 8 Separation of synthetic oligonucleotides by polymer-sieving CE. Separation was obtained in a neutral-coated capillary uisng a commercial polymer-sieving applications kit developed for analysis of oligonucleotides (Bio-Rad Laboratories; Hercules, CA).

polymers have low UV absorbance at short wavelengths, and can be replenished between analyses by pressure. Polymer solutions act by forming entangled networks above a certain polymer concentration, which mimics the sieving properties of cross-linked gels. Polymers used for sieving separations include alkylated celluloses, polyethylene oxide, and dextrans. Replaceable polymer systems have been developed for separation of SDS–protein complexes ranging form 10 to 200 kDa (Fig. 7). Replaceable polymer solutions have found wide use in nucleic acid separations, with applications including purity determination of synthetic oligonucleotides (Fig. 8), separation of DNA restriction fragments, DNA typing using microsatellite or short tandem repeat (STR) sequences, mutation analysis of amplified genetic sequences, and, most importantly, DNA sequence analysis. Because of their excellent precision, ease of use, and high resolution, polymer sieving systems have virtually replaced gel-filled capillaries for size-based separations by capillary electrophoresis.

VIII. ISOTACHOPHORESIS

As a separation techique, isotachophoresis (ITP; also known as displacement electrophoresis) resolves analytes as contiguous zones which migrate in order of mobility. The sample is injected into the capillary between a leading buffer (with ion mobility greater than that of all analytes) and a terminating buffer (with ion mobility less than that of all analytes). Zones migrate at equal velocity toward the detection point where they are detected as steps with zone length proportional to concentration. When UV-transparent spacers are added to the sample, ITP zones may appear as isolated peaks and the detector trace will resemble a CZE electropherogram. Isotachophoresis is rarely used as a separation method but is occasionally used transiently as an on-line preconcentration technique prior to CZE separation.

IX. MICELLAR ELECTROKINETIC CHROMATOGRAPHY

As the name implies, micellar electrokinetic chromatography (MEKC) is a chromatographic technique in which samples are separated by differential partitioning between two phases (Fig. 9). The technique is usually performed in uncoated capillaries under alkaline conditions to generate a high electroosmotic flow. The background electrolyte contains a surfactant at a concentration above its critical

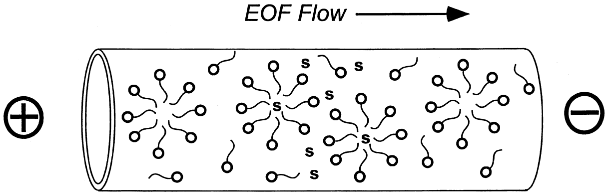


FIGURE 9 Schematic of an MEKC separation. Sample molecules are labeled "S."

micelle concentration (CMC); surfactant monomers are in equilibrium with micelles. The most widely used MEKC system employs sodium dodecylsulfate (SDS) as the surfactant. The sulfate groups of SDS are anionic, so both surfactant monomers and micelles have electrophoretic mobility counter to the direction of EOF. Sample molecules will be distributed between the bulk aqueous phase and the micellar phase, depending upon their hydrophobic character. Hydrophilic neutral species with no affinity for the micelle will remain in the aqueous phase and reach the detector in the time required for EOF to travel the effective length of the column. Hydrophobic neutral species will spend varying amounts of time in the micellar phase depending on their hydrophobicity, and their migration will therefore be retarded by the anodically moving micelles. Charged species will display more complex behavior, and their migration will include contributions from electrophoretic mobility and electrostatic interaction with the micelles in addition to hydrophobic partitioning. The selectivity of MEKC can be expanded with the introduction of chiral selectors to the system. MEKC is used almost exclusively for small molecules such as drugs and metabolites, pesticides, herbicides, vitamins, etc.

X. CAPILLARY ELECTROKINETIC CHROMATOGRAPHY

Like MEKC, capillary electrokinetic chromatography (CEC) is a chromatographic technique performed with CE instrumentation. It employs fused silica capillaries packed with 1.5- to 5-μm microparticulate porous silica beads, usually derivatized with a hydrophobic ligand such as C18. Mobile phases are similar to those used for conventional reversed-phase HPLC (e.g., mixtures of aqueous buffers and an organic modifier such as acetonitrile). The silica surface of the derivatized beads has sufficient densities of ionized silanol groups to generate a high electroosmotic flow when a voltage is applied to the system, pumping mobile phase through the column. Since EOF is plug-like rather than laminar in nature, efficiencies in CEC can be much higher than in HPLC. In contrast to pressure-driven capillary liquid chromatography, which requires pressures of several thousand psi to pump mobile phase through the column, CEC generates no column backpressure. Like MEKC, CEC is used primarily for small molecules.

SEE ALSO THE FOLLOWING ARTICLES

ELECTROPHORESIS • LIQUID CHROMATOGRAPHY • MASS SPECTROMETRY

BIBLIOGRAPHY

Camilleri, P., ed. (1993). "Capillary Electrophoresis: Theory and Practice," CRC Press, Boca Raton, FL.

Deyl, Z., and Struzinsky, R. (1991). *J. Chromatogr.* **569,** 63.

Grossman, P. D., and Colburn, J. C., eds. (1992). "Capillary Electrophoresis: Theory and Practice," Academic Press, San Diego, CA.

Guzman, N. A. (1993). "Capillary Electrophoresis Technology," Dekker, New York.

Kuhr, W. G. (1990). *Anal. Chem.* **62,** 403R.

Kuhr, W. G., and Monnig, C. A. (1992). *Anal. Chem.* **64,** 389R.

Landers, J. P., ed. (1996). "Handbook of Capillary Electrophoresis," 2nd ed., CRC Press, Boca Raton, FL.

Li, S. F. Y. (1992). "Capillary Electrophoresis: Principles, Practice, and Applications," Vol. 52, Elsevier Science, Amsterdam.

Lunte, S. M., and Radzik, D. M., eds. (1996). "Pharmaceutical and Biomedical Applications of Capillary Electrophoresis: Progress in Pharmaceutical and Biomedical Analysis," Vol., Elsevier Science, Oxford.

Monnig, C. A., and Kennedy, R. T. (1994). *Anal. Chem.* **66,** 280R.

Wehr, T., Rodriguez-Díaz, R., and Zhu, M. (1998). "Capillary Electrophoresis of Proteins," Dekker, New York.

Weinberger, R. (1992). "Practical Capillary Electrophoresis," Academic Press, Boston, MA.

Vindevogel, V. (1992). "Introduction to Micellar Electrokinetic Chromatography," Hüthig, Heidelberg.

Electrochemistry

Donald T. Sawyer
Texas A&M University

I. Fundamentals
II. Electrode Potentials and Potentiometry
III. Controlled-Potential Electrolysis and Voltammetry
IV. Electron-Transfer Processes
V. Electrochemical Characterization of Molecules
VI. Industrial Electrosynthesis
VII. Batteries and Fuel Cells
VIII. Corrosion; Cathodic Protection

GLOSSARY

Anode Electrode at which a chemical oxidation occurs with removal of electrons; the negative terminal for spontaneous cell reactions (batteries, galvanic cells, and fuel cells); the positive terminal for imposed cell reactions (electrolysis cells and voltammetric cells).

Cathode Electrode at which a chemical reduction occurs via introduction of electrons; the positive terminal for spontaneous cell reactions (batteries, galvanic cells, and fuel cells); the negative terminal for imposed cell reactions (electrolysis cells and voltammetric cells).

Cell Container that includes the electrodes (anode and cathode) connected via an electrolyte solution. Physical contact between the anode and the cathode and mixing of electrolysis products from the two electrodes often is prevented by the use of separators (membranes or inert materials).

Electroactive species Molecules that are transformed at the anode and cathode.

Electrode Conducting material (usually metal or carbon) that is immersed in an electrolyte solution and is a source or sink for electrons.

Electrolysis Passage of ionic current through an electrolyte solution with reduction of an electroactive species at the cathode and oxidation of an electroactive species at the anode.

Electrolyte Solution that conducts ionic current via ionic transport (ions of dissolved salts, mineral acids, or strong bases).

Indicator electrode Electrode that responds to the activity (thermodynamic effective concentration) of a solution species in potentiometric and voltammetric measurements.

Reference electrode An electrochemical half-cell with a stable and known potential. The most common is the saturated calomel electrode (SCE, +0.242 V vs NHE). The primary standard reference electrode is the normal hydrogen electrode (NHE, 0.0000 V by definition).

THE TECHNOLOGY for the interconversion of chemical energy and electrical energy has been utilized since the mid-19th century. This conversion is accomplished by ionic-current flow in an electrolyte solution between two electrodes connected to each other via an external circuit with an electrical load or current source. Batteries, fuel cells, and corrosion processes convert the energy of chemical reactions into electrical energy. Electrolysis, electroplating, and some forms of electroanalysis reverse the direction of conversion, using electrical energy to produce a net chemical change. The basic principles and quantitative relationships (voltage, current, charge conductance, capacitance, and concentration) for electrochemical phenomena were empirically elucidated by Michael Faraday and other European scientists before the discovery of the electron (J. J. Thompson, 1893) and the development of chemical thermodynamics (G. N. Lewis, 1923). Building on this foundation, the utilization of electrochemical phenomena for thermodynamic characterization and analysis of molecules and ions (electroanalytical chemistry) began at the beginning of this century [potentiometry (1920) and polarography (1930)]. Relationships that describe the techniques of potentiometry and polarography derive directly from solution thermodynamics. In the case of polarography, there is a further dependence on the diffusion of ionic species in solution. The latter is the basis of conductivity measurements, another area that traces its origin to the 19th century. These quantitative relationships make it possible to apply electrochemistry to the detailed characterization of chemical species and processes in the solution phase.

Electrochemistry is the science of electron transfer across a solution/electrode interface. At the cathode, electrons (from the electrode) are transformed within the interface via reaction with ions or molecules to produce reduced molecules or ions (e.g., $H_3^+O + e^- \rightarrow H\cdot + H_2O$; $H_2O + e^- \rightarrow H\cdot + HO^-$; $\cdot O_2\cdot + e^- \rightarrow O_2^-\cdot$; $Cu^{II}(bpy)_2^{2+} + e^- \rightarrow Cu^{I}(bpy)_2^+$; $Fe^{III}Cl_3 + e^- \rightarrow Fe^{II}Cl_3^-$). [Note: Although the traditional formulation of the hydronium ion (H_3O^+) is pervasive in the chemical literature, the positive charge is equally distributed among the three hydrogens, which prompts the formulation used here (H_3^+O).] At the anode, molecules or ions (from the solution) are transformed within the interface to produce electrons (at the electrode surface) and oxidized ions and molecules (e.g., $2\,H_2O \rightarrow H_3^+O + HO\cdot + e^-$; $Fe^{II}Cl_3^- \rightarrow Fe^{III}Cl_3 + e^-$). The resultant electrons move from the anode through the wires of the external circuit to the cathode as electronic current (amperes; coulombs per second). Within the solution phase the current is carried by the ions of the supporting electrolyte (positive ions toward the cathode and negative ions toward the anode). The limitation of ionic current in the solution phase (between the anode and the cathode), which is the defining difference for electrochemistry and electronics, is due to the incompatibility of electrons and electrolyte solutions.

During the past four decades, the dynamics and mechanisms of electron-transfer processes have been studied via the application of transition-state theory to the kinetics for electrochemical processes. As a result, the kinetics of both the electron-transfer processes (from solid electrode to the solution species) as well as the pre- and post-electron-transfer homogeneous processes can be characterized quantitatively.

By the use of various transient methods, electrochemistry has found extensive new applications for the study of chemical reactions and adsorption phenomena. Thus, a combination of thermodynamic and kinetic measurements can be utilized to characterize the chemistry of heterogeneous electron-transfer reactions. Furthermore, heterogeneous adsorption processes (liquid-solid) have been the subject of intense investigations. The mechanisms of metal-ion complexation reactions also have been ascertained through the use of various electrochemical impulse techniques.

The so-called Renaissance of electrochemistry has come about through a combination of modern electronic instrumentation and the development of a more molecular-based theory implemented with the data processing and computational power of computers. Within the area of physical chemistry, numerous thermodynamic studies of unstable reaction intermediates have made use of modern electrochemistry. In addition, extensive studies of the kinetics of electron-transfer processes in aqueous and nonaqueous media have been accomplished. The electrochemical characterization of adsorption phenomena has been of immense benefit to the understanding of catalytic processes.

Some of the most exciting applications of electrochemistry have occurred in the areas of organic and inorganic chemistry and of biochemistry. The applications have ranged from mechanistic studies to the synthesis of unstable or exotic species. The control of an oxidation or reduction process through electrochemistry is much more precise than is possible with chemical reactants. Within the area of inorganic chemistry, electrochemistry has been especially useful for the determination of formulas of coordination complexes and the electron-transfer stoichiometry of new organometallic compounds. Electrochemical synthesis is increasingly important to the field of organometallic chemistry.

During the past 50 years, numerous exciting extensions of electrochemistry to the field of analytical chemistry have occurred. A series of selective-ion potentiometric electrodes have been developed, such that most of the common ionic species can be quantitatively monitored in

aqueous solution. A highly effective electrolytic moisture analyzer provides continuous online assays for water in gases. Another practical development has been the voltammetric membrane electrode for dioxygen ($\cdot O_2\cdot$), which responds linearly to the partial pressure of $\cdot O_2\cdot$, either in the gas phase or in solution. The use of an immobilized enzyme (glucose oxidase) on an electrode sensor to assay glucose in blood is another extension of electrochemistry to practical analysis.

I. FUNDAMENTALS

Reductive electrochemistry involves the transfer of electrons from the electrode surface (cathode) into the double-layer interface. Within the latter, the electrons react with the most electrophilic component of the interface; e.g., $H_3^+O(aq)$ in acidic aqueous solutions and $H_2O(aq)$ in basic aqueous solutions. From electrochemistry the respective standard reduction potentials in water are

$$e^- + H_3^+O \longrightarrow [H\cdot] + H_2O \quad E^\circ, -2.10 \text{ V vs NHE} \quad (1)$$

$$e^- + H_2O \longrightarrow [H\cdot] + HO^- \quad E^\circ, -2.93 \text{ V vs NHE} \quad (2)$$

and in acetonitrile are

$$e^- + H_3^+O \longrightarrow [H\cdot] + H_2O \quad E^{\circ\prime}, -1.58 \text{ V vs NHE} \quad (3)$$

$$e^- + H_2O \longrightarrow [H\cdot] + HO^- \quad E^{\circ\prime}, -3.90 \text{ V vs NHE.} \quad (4)$$

A. Electron-Transfer Dynamics (Kinetics and Thermodynamics)

The rate of electron transfer at an electrode/solution interface for the direct reduction of an oxidized species (*ox*) to its reduced state (*red*) [e.g., $Cu^{II}(bpy)_2^{2+} + e^- \rightarrow Cu^{I}(bpy)_2^{+}$],

$$ox + ne^- \underset{k_a}{\overset{k_c}{\rightleftharpoons}} red, \quad (5)$$

is a function of the concentration of the oxidized species and its heterogeneous electron-transfer rate constant (k_c; cathodic process). The latter is a function of the potential difference across the electrode/solution interface (ΔE), which is directly proportional to the activation energy for reduction ($-\Delta G_c^\ddagger$). Only a fraction of ΔE is effective for accelerating the rate of reduction, which is represented by a symmetry parameter, α [transfer coefficient, $0.0 < \alpha < 1.0$ (usually about 0.5)],

$$\begin{aligned}\left(\Delta G_c^\ddagger\right)_{\text{total}} &= \left(\Delta G_c^\ddagger\right)_{\Delta E=0} + \left(\Delta G_c^\ddagger\right)_{\Delta E} \\ &= \left(\Delta G_c^\ddagger\right)_{\Delta E=0} + \alpha n_a \Delta E F, \end{aligned} \quad (6)$$

where n_a = number of electrons in the rate limiting step; usually one.

The flux of *ox* that is reduced at the electrode (v_c) has the dimensions of moles per unit area (cm^2) per second, and

$$v_c = d(C_{ox})_o / A\,dt = k_c(C_{ox})_o = j_c/nF), \quad (7)$$

where $(C_{ox})_o$ is concentration (mol/cm^3) at the electrode surface, A is area of the electrode (cm^2), and j_c is net cathodic (reductive) current density (A/cm^2).

At equilibrium ($\Delta E = 0$),

$$(v_c)_{\Delta E=0} = (C_{ox})_o[\kappa kT/h] \exp\left[-\left(\Delta G_c^\ddagger\right)_{\Delta E=0}/RT\right], \quad (8)$$

where κ is the transmission coefficient within the Activated Complex Theory, k is the Boltzmann constant, and h is the Planck constant, which, in combination with Eq. (7), gives an expression for the heterogeneous electron-transfer rate constant when ΔE is zero,

$$(k_c)_{\Delta E=0} = [\kappa kT/h] \exp\left[-\left(\Delta G_c^\ddagger\right)_{\Delta E=0}/RT\right]. \quad (9)$$

In turn, combination of these equations gives expressions for the flux of *ox* reduction (v_c) and cathodic current density (j_c) for a potential difference ΔE.

$$\begin{aligned} v_c &= (C_{ox})_o[\kappa kT/h] \exp\left[-\left(\Delta G_c^\ddagger\right)_{\Delta E=0}/RT\right] \\ &\quad \times \exp(-\alpha n_a \Delta E F/RT) \\ &= (C_{ox})_o (k_c)_{\Delta E=0} \exp(-\alpha n_a \Delta E F/RT); \end{aligned} \quad (10)$$

$$j_c = (C_{ox})_o\, nF(k_c)_{\Delta E=0} \exp(-\alpha n_a \Delta E F/RT). \quad (11)$$

The expressions for the reverse anodic process (oxidation) of Eq. (5) ($red \rightarrow ox + n\,e^-$; k_a) follow from similar arguments:

$$v_a = (C_{red})_o (k_a)_{\Delta E=0} \exp[(1-\alpha)n_a\, \Delta E F/RT], \quad (12)$$

$$j_a = (C_{red})_o\, nF(k_a)_{\Delta E=0} \exp[(1-\alpha)n_a\, \Delta E F/RT]. \quad (13)$$

When the cathodic current density (j_c) is equal to the anodic current density (j_a), the net current flow across the electrode/solution interface is zero and the net flux of *ox* and *red* is zero. For this unique condition (zero net current) the current densities represent the equilibrium exchange current density (j_o),

$$j_c = j_a = j_o, \quad (14)$$

which is associated with the equilibrium potential difference, ΔE_e. Thus,

$$j_o = j_c = (C_{ox})_o\, nF(k_c)_{\Delta E=0} \exp(-\alpha \Delta E_e F/RT) \quad (15)$$

$$j_o = j_a = (C_{red})_o\, nF(k_a)_{\Delta E=0} \exp[(1-\alpha)\Delta E_e F/RT]. \quad (16)$$

The difference between ΔE and ΔE_e is the activation overpotential (η),

$$\eta = \Delta E - \Delta E_e. \quad (17)$$

These relationships can be combined to give an expression for the net current density (j), which by definition is equal to $j_a - j_c$ in terms of the activation overpotential (substitute $\eta + E_e = \Delta E$), which is referred to as the Butler-Volmer equation

$$j = j_a - j_c = j_o\{\exp[(1-\alpha)n_a F\eta/RT] - \exp[-\alpha n_a F\eta/RT]\}. \quad (18)$$

Both k_c and k_a vary exponentially with potential difference:

$$k_c = (k_c)_{\Delta E=0}\exp[-\alpha n_a \Delta EF/RT], \quad (19)$$

$$k_a = (k_a)_{\Delta E=0}\exp[(1-\alpha)n_a\, \Delta EF/RT]. \quad (20)$$

At equilibrium ($j = 0$) and for the special case when $(C_{ox})_o = (C_{red})_o$,

$$(k_c)_{\Delta E=0} = (k_a)_{\Delta E=0} = k_s. \quad (21)$$

This simple rate constant (k_s) is the defined value of k_c and k_a at the formal potential, $\Delta E^{\circ\prime}$. Then,

$$(k_c)_{\Delta E=0} = k_s \exp[-\alpha n_a F(\Delta E - \Delta E^{\circ\prime})/RT] \quad (22)$$

$$(k_a)_{\Delta E=0} = k_s \exp[(1-\alpha)n_a F(\Delta E - \Delta E^{\circ\prime})/RT] \quad (23)$$

$$j = nFk_s\{(C_{red})_o \exp[(1-\alpha)n_a F \times (\Delta E - \Delta E^{\circ\prime})/RT] - (C_{ox})_o \exp[-\alpha n_a F(\Delta E - \Delta E^{\circ\prime})/RT]\}. \quad (24)$$

Redox couples in aqueous solutions at room temperature have k_s values that range from about 1.0 cm s^{-1} down to essentially zero (<10^{-10} cm s^{-1}); e.g., $k_s = 0.1$ cm s^{-1} for $Fe^{III}(CN)_6^{3-}/Fe^{II}(CN)_6^{4-}$ and 10^{-5} cm s^{-1} for $Cr^{III}(OH_2)_6^{3+}/Cr^{II}(OH_2)_6^{2+}$.

At the equilibrium potential, $\Delta E = \Delta E_e$, $j = 0$, $C^o_{ox} = C^b_{ox}$, and $C^o_{red} = C^b_{red}$, where C^o denotes the concentration of a species at the electrode surface, whereas C^b its bulk concentration. Then

$$\frac{C^b_{bx}}{C^b_{red}} = \exp[-nF(\Delta E - \Delta E^{\circ\prime})/RT] \quad (25)$$

and

$$\Delta E_e = \Delta E^{\circ\prime} + (RT/nF)\ln\frac{C^b_{ox}}{C^b_{red}} = \Delta E^{\circ\prime} + (0.05915/n)\log([ox]/[red]) \quad (26)$$

(Nernst equation for a half reaction).

Also, at the equilibrium potential (ΔE_e),

$$j_o = nFk_s\left(C^b_{ox}\right)^{(1-\alpha)}\left(C^b_{red}\right)^{\alpha} = nFk_s C \quad (\text{when } C_{ox} = C_{red}). \quad (27)$$

More modern models and treatments for the dynamics of electrochemical electrontransfer (of the Marcus/Hush/Gerischer type) are discussed in Bard and Faulkner (2001).

II. ELECTRODE POTENTIALS AND POTENTIOMETRY

Use of the potential of a galvanic cell to measure the concentration of an electroactive species developed later than a number of other electrochemical methods. In part, this was because a rational relation between electrode potential and the concentration of an electroactive species required the development of thermodynamics, and, in particular, its application to electrochemical phenomena. The work of J. Willard Gibbs in the 1870s provided the foundation for the Nernst equation. The latter provides a quantitative relationship between potential and the ratio of effective thermodynamic concentrations [activities] for a redox couple ([*ox*]/[*red*]) and is the basis for potentiometry and potentiometric titrations. The utility of potentiometric measurements for the characterization of ionic solutions was established with the invention of the glass electrode in 1909 for a selective potentiometric response to hydronium-ion concentrations. Another milestone in the development of potentiometric measurements was the introduction of the hydrogen electrode for the measurement of hydronium-ion concentrations; one of many important contributions by Professor Joel Hildebrand. Subsequent development of special glass formulations has made possible electrodes that are selective to different monovalent cations. The idea is so attractive that intense effort has led to the development of electrodes that are selective for many cations and anions, as well as several gas- and bioselective electrodes. The use of these electrodes and the potentiometric measurement of pH continue to be among the most important applications of electrochemistry.

A. Principles and Fundamental Relations

Potentiometric measurements are based on thermodynamic relationships and, more particularly, the Nernst equation which relates potential to the concentration of electroactive species. For our purposes, it is most convenient to consider the redox process that occurs at a single electrode, although two electrodes are always essential for an electrochemical cell. However, by considering each electrode individually, the two electrode processes are easily combined to obtain the entire cell process. Furthermore, confusion can be minimized if the half reactions for electrode processes are written in a consistent manner. Here, these are always reduction processes with the oxidized species reduced by n electrons to give a reduced species,

$$ox + ne^- \rightarrow red. \quad (28)$$

For such a half reaction the free energy is given by the relation

$$\Delta G = \Delta G^\circ + RT \ln \frac{[red]}{[ox]}, \quad (29)$$

where $-\Delta G$ indicates the tendency for the reaction to go to the right; R is the gas constant and in the units appropriate for electrochemistry has a value of 8.317 J mol^{-1} K^{-1}; T is the temperature of the system in K; and the logarithmic terms in the bracketed expression represent the activities (effective concentrations) of the electroactive pair at the electrode surface. The free energy of this half reaction is related to the electrode potential, E, by the expression

$$-\Delta G = nFE; \qquad -\Delta G^\circ = nFE^\circ. \quad (30)$$

The quantity ΔG° is the free energy of the half reaction when the activities of the reactant and product have values of unity and is directly proportional to the standard half-cell potential for the reaction as written. It also is a measure of the equilibrium constant for the half reaction, assuming the activity of electrons is unity,

$$-\Delta G^\circ = RT \ln K. \quad (31)$$

An extensive summary of E° values is presented in the compilation by Bard, Parsons, and Jordan (1985); the most important are tabulated in Sawyer, Sobkowiak, and Roberts (1995). Standard potentials are thermodynamic quantities that usually are evaluated via caloriometry for a cell reaction (e.g., $2\ H_2 + \cdot O_2 \cdot \rightarrow 2H_2O$; $E^\circ_{cell} = [E^\circ_{O_2} - E^\circ_{H^+/H_2}]$) and the relationship of Eq. (30).

When Eqs. (29) and (30) are combined, the resulting Nernst expression relates the half-cell potential to the effective concentrations (activities) of the redox couple,

$$E = E^\circ - \frac{RT}{nF} \ln \frac{[red]}{[ox]} = E^\circ + \frac{RT}{nF} \ln \frac{[ox]}{[red]}. \quad (32)$$

The activity of a species is indicated as the symbol of the species enclosed in a bracket. This quantity is equal to the concentration of the species times a mean activity coefficient,

$$[M^{a+}] = a_{M^{a+}} = \gamma_\pm C_{M^{a+}}. \quad (33)$$

Although there is no straightforward and convenient method for evaluating activity coefficients for individual ions, the Debye-Huckel relationship permits an evaluation of the mean activity coefficient ($\gamma_\pm$) for ions at low concentrations (usually below 0.01 M),

$$\log \gamma_\pm = -0.509 z^2 \frac{\sqrt{\mu}}{1 + \sqrt{\mu}}, \quad (34)$$

where z is the charge on the ion and μ is the ionic strength

$$\mu = \left(\frac{1}{2}\right) \sum C_i z_i^2. \quad (35)$$

More rigorous methods for the calculation of aqueous activity coefficients are available.

The reaction of an electrochemical cell always involves a combination of two redox half reactions such that one species oxidizes a second species to give the respective redox products. Thus, the overall cell reaction can be expressed by a balanced chemical equation

$$a\ ox_1 + b\ red_2 \rightarrow c\ red_1 + d\ ox_2, \qquad K_{equil}. \quad (36)$$

However, electrochemical cells are most conveniently considered as two individual half reactions, whereby each is written as a reduction in the form indicated by Eqs. (28)–(32). When this is done and values of the appropriate quantities are inserted, a potential can be calculated for each half-cell electrode system. Then that half-cell reaction with the more positive potential will be the positive terminal in a galvanic cell, and the electromotive force of that cell will be represented by the algebraic difference between the potential of the more positive half-cell and the potential of the less positive half-cell,

$$E_{cell} = E_{(\text{more positive})} - E_{(\text{less positive})} = E_1 - E_2. \quad (37)$$

Insertion of the appropriate forms of Eq. (32) into Eq. (37) gives an overall expression for the cell potential,

$$E_{cell} = E_1^\circ - E_2^\circ + \frac{RT}{nF} \ln \frac{[ox_1]^a [red_2]^b}{[ox_2]^d [red_1]^c}. \quad (38)$$

The equilibrium constant for the chemical reaction expressed by Eq. (36) is related to the difference of the standard half-cell potentials by the relation

$$\ln K_{equil} = (nF/RT)(E_1^\circ - E_2^\circ). \quad (39)$$

To apply potentiometric measurements to the determination of the concentration of electroactive species, a number of conditions have to be met. The basic measurement system must include an indicator electrode, which is capable of monitoring the activity of the species of interest, and a reference electrode, which gives a constant, known half-cell potential to which the indicator electrode potential can be referred. The voltage resulting from the combination of these two electrodes must be measured in a manner that minimizes the amount of current drawn by the measuring system. For low-impedance electrode systems, a conventional potentiometer is satisfactory. However, electrochemical measurements with high-impedance electrode systems, and in particular the glass-membrane electrode, require the use of an exceedingly high-input-impedance measuring instrument (usually an electrometer amplifier with a current drain of less than 10^{-12} A). Because of the logarithmic nature of the Nernst equation, the measuring instrumentation must have considerable sensitivity. For example, a one-electron half reaction of 25°C gives a voltage change of 59.1 mV for a 10-fold change

in the concentration of the electroactive species. Another important point is that the potential response is directly dependent on the temperature of the measuring system. Thus, if the correct temperature is not used in the Nernst expression, large absolute errors can be introduced in the measurement of the activity for an electroactive species.

B. Electrode Systems

The indicator electrodes for potentiometric measurements traditionally have been categorized into four separate classes. "First-class" electrodes consist of a metal immersed in a solution that contains the metal ion. These electrode systems provide a direct response to the ion or species to be measured,

$$E = E^\circ + \frac{RT}{nF}\ln\,[M^{n+}]; \qquad M^{n+} + ne^- \rightleftharpoons M\,(\mathrm{s}). \tag{40}$$

Therefore, the primary electrode reaction includes the sensed species. Such electrodes give a direct response according to the Nernst equation for the logarithm of the activity of the species.

Electrodes classified as "second-class" electrode systems are those in which the electrode is in direct contact with a slightly soluble salt of the electroactive species such that the potentiometric response is indicative of the concentration of the inactive anion species. Thus, the silver/silver-chloride electrode system, which is representative of this class of electrodes, gives a potential response that is directly related to the logarithm of the chloride ion activity,

$$E = E^\circ - \frac{RT}{nF}\ln\,[\mathrm{Cl}^-]; \quad \mathrm{AgCl\,(s)} + e^- \rightleftharpoons \mathrm{Ag\,(s)} + \mathrm{Cl}^-, \tag{41}$$

even though it is not the electroactive species. This is true because the chloride-ion concentration, through the solubility product, controls the activity of the silver ion, which is measured directly by the potentiometric silver-electrode system.

Because any potentiometric electrode system ultimately must have a redox couple (or an ion-exchange process in the case of membrane electrodes) for a meaningful response, the most common form of potentiometric electrode systems involves oxidation-reduction processes. Hence, to monitor the activity of ferric ion [iron(III)], an excess of ferrous ion [iron(II)] is added such that the concentration of this species remains constant to give a direct Nernstian response for the activity of iron(III). For such redox couples the most common electrode system has been the platinum electrode. This tradition has come about primarily because of the historic belief that the platinum electrode is totally inert and involves only the pure metal as a surface. However, during the past 50 years, it has become evident that platinum electrodes are not as inert as long believed and that their potentiometric response is frequently dependent on the history of the surface and the extent of its activation. The evidence is convincing that platinum electrodes, and, in fact, all metal electrodes, are covered with an oxide film that changes its characteristics with time. Nonetheless, the platinum electrode continues to enjoy wide popularity as an "inert" indicator of redox reactions and of the activities of the ions involved in such reactions.

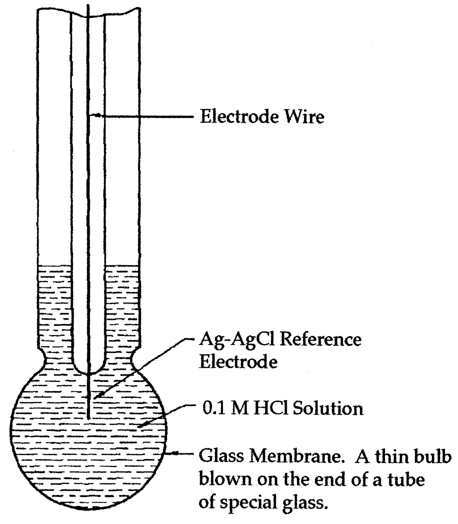
For those redox couples that involve a metal ion plus the metal, the logical electrode system is the metal itself. In other words, if the measured quantity is to be cupric ion [copper (II)], a practical indicator electrode is a piece of copper metal. All second-class electrodes involve an active metal in combination with an insoluble compound or salt. Thus, the silver/silver-chloride electrode actually is a silver/silver-ion electrode system that incorporates the means to control the silver-ion concentration through the chloride-ion concentration [Eq. (41)].

A number of the most common potentiometric electrode systems and their applications are summarized in Table I. One of the most important and extensively used indicator electrode systems is the glass-membrane electrode that is used to monitor hydronium-ion activity. Although developed in 1909, it did not become popular until reliable electrometer amplifiers were developed in the 1930s (modern pH meters use high-input-impedance digital voltmeters). Figure 1 gives a schematic representation of this electrode and indicates that the primary electrode system is a silver/silver-chloride (or mercury/mercurous-chloride) electrode in contact with a known and fixed concentration of hydrochloric acid (usually about 0.1 M). When

TABLE I Redox Potentials ($E^{\circ\prime}$) for the $M^I(OH_2)_2^+/M$ and $M^IOH/M, HO^-$ Couples of Cu, Ag, and Au in H_2O and in MeCN (0.1 M Tetraethylammonium Perchlorate)

	M	E°, V vs NHE[a] $M^I(solv)_n^+/M$	$M^IOH/M, HO^-$
A. H_2O	Cu	+0.52	−0.36
	Ag	+0.80	+0.34
	Au	+1.7	
	$H_2^+O\cdot/H_2O$; $HO\cdot/HO^-$(GC)	+2.72	+1.89
B. MeCN	Cu	+0.19	−0.79
	Ag	+0.54	−0.30
	Au	+1.58	−0.19
	$H_2^+O\cdot(MeCN)/H_2O$; $HO\cdot/HO^-$(GC)	+3.2	+0.92

[a] SCE = +0.24 V vs NHE.



(a)

Ag, AgCl	0.1 M HCl	Glass membrane	Unknown solution	satd. KCl	Hg_2Cl_2, Hg

(b)

FIGURE 1 Glass electrode (a) and its cell schematic (b) in association with a reference electrode.

the outside surface of the glass membrane is exposed to an ionic solution, a response for the hydronium-ion activity $[H_3^+O]$ is obtained that follows the Nernst expression. Although there has been considerable debate about the mechanism of response for the glass electrode, the current thinking invokes an ion-exchange process that involves the hydroxyl groups on the surface of the glass. Thus, the population of protons on the outside surface of the membrane affects the population on the inside of the membrane, which generates a membrane potential that is indicated by the silver/silver chloride. At one time there was a belief that hydronium ions actually penetrated the glass membrane from the outside to the inside. However, experiments with labeled systems establish that this is not true. Further support for the ion-exchange mechanism is provided by the realization that glass electrodes are not specific for hydronium ions, but only give a selective response. In other words, other metal ions, in particular those of sodium and lithium, cause a response from glass electrodes through an equivalent ion-exchange process.

The unwanted response of glass electrodes to metal ions, particularly alkali metal ions, has prompted the development of specialized glass membranes that have an enhanced selective response for monovalent cations. A parallel approach has been used in the development of ion-exchange membranes (prepared from organic, polymer-based, ion-exchange resins) that give a selective response to specific cations and anions. The main advantages of the membranes are resistance to high acid and alkali concentrations and high conductivity. The membranes, however, show little selectivity between ions. A related form of the membrane electrode is the inorganic-crystal electrode. An example is the lanthanum-fluoride (LaF_3) electrode that senses fluoride-ion concentrations, which has become the standard sensor for fluoride determinations in water analysis. It represents a combination of a membrane electrode with ion-exchange characteristics and an incomplete form of a second-class electrode in which lanthanum fluoride is the insoluble material that responds to the free fluoride-ion concentration in the test sample. Other examples of this form of electrode include AgX/Ag_2S and MS/Ag_2S electrodes, which give response to monovalent anions and divalent cations, respectively.

In contrast to solid-membrane electrodes, liquid-membrane electrodes can extract counterions from the solution phase into the membrane phase. Selectivity is provided by the charged nature of the membrane carriers and arises from the competitive degree of extractability of various counterions. Totally liquid systems can be employed but are impractical. Instead, a porous support or an inert polymer support is used in most commercial electrodes.

Gas-selective electrodes are a particularly important application of the glass electrode. For example, the carbon-dioxide electrode is a self-contained system with a glass electrode and a concentric silver/silver-chloride electrode enclosed by a CO_2 permeable membrane. The latter holds a thin film of bicarbonate solution in contact with the glass membrane, which provides a junction to the silver/silver-chloride reference electrode. The electrode has found extensive application in monitoring CO_2 levels in blood and probably will find increasing application in other systems that require continuous measurement of CO_2 partial pressures. The electrode response is based on the reaction

$$CO_2 + 2\,H_2O \rightleftharpoons H_3^+O + HOC(O)O^- \qquad (42)$$

such that changes in the partial pressure of carbon dioxide cause an attendant change in the concentration of hydronium ion. [Note: Although the coventional formulation of bicarbonate ion (HCO_3^-) is common, the formulation used here ($HOC(O)O^-$) provides a better sense of its structure and chemistry.] Thus, with a fixed concentration of bicarbonate the electrode provides a direct potentiometric response to the partial pressure of carbon dioxide. Other gas-monitoring electrode systems should be possible that are based on similar processes. For example, an ammonia (NH_3) electrode might well be developed with the converse of the reactions indicated for the CO_2 electrode. Thus, an ammonium-ion (NH_4^+) electrolyte would be used

such that changes in pH would be proportional to changes in the partial pressure of NH_3. These electrodes as well as SO_x-, NO_x-, HF-, H_2S-, and HCN-sensing electrodes are commercially available.

The gas-sensing electrodes also are used for the potentiometric measurement of biologically important species. An enzyme is immobilized at or near the gas probe. The gas sensor measures the amount of characteristic gas produced by the reaction of the analyzed substance with the enzyme. For example, an enzyme electrode for urea [$NH_2C(O)NH_2$] determination is constructed by the immobilization of *urease* onto the surface of an ammonia-selective electrode. When the electrode is inserted into a solution that contains urea, the enzyme catalyzes its conversion to ammonia,

$$NH_2C(O)NH_2 + H_2O \xrightarrow{\text{urease}} 2\,NH_3 + CO_2. \quad (43)$$

The generated ammonia diffuses through the gas-permeable membrane and is detected by the probe of the electrode. A steady-state signal is reached when the rate of NH_3 diffusion from the electrode equals the rate of its generation.

In general, a necessary part of a potentiometric measurement is the coupling of a reference electrode to the indicating electrode. The ideal reference electrode has a number of important characteristics: (1) a reproducible potential, (2) a low-temperature coefficient, (3) the capacity to remain unpolarized when small currents are drawn, and (4) inertness to the sample solution. If the reference electrode must be prepared in the laboratory, a convenient and reproducible system is desirable.

Although the standard half-cell reactions are all referenced to the standard hydrogen electrode (NHE or SHE), this is an exceedingly awkward reference electrode. It has been selected because its potential falls in the middle of those for the most common half reactions in water and because with rigorous care highly reproducible potentials can be duplicated by equally careful workers in other laboratories. Furthermore, because it consists of a platinized platinum electrode over which hydrogen gas is passed in combination with a known activity of hydronium ion, it can be combined in a number of cases with other half-cells without a liquid junction. These three factors undoubtedly justify its selection as the ultimate reference electrode for fundamental measurements. The electrode reaction is a typical redox half reaction which includes the oxidized and reduced forms of hydrogen. By controlling the hydrogen partial pressure at a fixed level, this becomes an indicating electrode for hydronium-ion activity.

For most potentiometric measurements, either the saturated calomel reference electrode or the silver/silver-chloride reference electrode is used. These electrodes can be made compact, are easily produced, and provide reference potentials that do not vary more than a few millivolts. The silver/silver chloride electrode also finds application in nonaqueous titrations, although some solvents cause the silver-chloride film to become soluble. Aqueous reference electrodes are as convenient for nonaqueous systems as any of the prototypes that have been developed to date. When there is a need to rigorously exclude water, double-salt bridges (aqueous/nonaqueous) are a convenient solution.

For measurement of redox couples, a frequently overlooked but convenient reference electrode is a conventional glass pH electrode (assuming the sample solution system contains a constant level of acidity). Such an electrode provides an extremely inert and stable reference potential that is completely indifferent to most redox species. However, the glass electrode requires the use of an electrometer amplifier such as that contained in pH meters.

C. Applications of Potentiometry

Although all potentiometric measurements (except those involving membrane electrodes) ultimately are based on a redox couple, the method can be applied to oxidation-reduction processes, acid-base processes, precipitation processes, and metal-ion complexation processes. Measurements that involve a component of a redox couple require that either the oxidized or reduced conjugate of the species to be measured be maintained at a constant and known activity at the electrode. If the goal is to measure the activity of silver ion in a solution, then a silver wire coupled to the appropriate reference electrodes makes an ideal potentiometric system. Likewise, if the goal is to monitor iron(III) concentrations with a platinum electrode, a known concentration of iron(II) must be present in the sample solution such that potential changes are only dependent on the iron(II) concentration.

Table I summarizes a number of redox couples that are well behaved in aqueous solutions and provide a means for monitoring the indicated species by potentiometric measurements. This can be either in the form of monitoring a titration or as a direct absolute measurement of activity. Although the tabulations of standard potentials infer that the listing should be much more comprehensive, most of the couples tabulated are not well behaved in an electrochemical sense and do not provide a Nerstian response under normal laboratory conditions. The vast majority of the data tabulated is based on calorimetric data.

The major application of the potentiometric method is for acid-base measurements in both aqueous and nonaqueous solvent systems. Although the glass electrode is universally the most common indicating electrode system for such measurements, many other electrodes have been developed. However, except in extremely specialized

circumstances, none of these provides the reliability and precision that is afforded by the glass electrode. In the absence of interfering substances, the quinhydrone electrode (an equimolar combination of quinone and hydroquinone with a gold-foil electrode) provides a simple monitoring system for measurements up to pH 8. However, the presence of oxidizing or strongly reducing ions in the sample system will interfere, as is true for the hydrogen gas electrode and most other systems that are an alternative to the glass electrode.

Potentiometric redox measurements are often performed in nonaqueous or mixed-solvent media. For such solvents, various potentiometric sensors have been developed, which, under rigorously controlled conditions, give a Nernstian response over a wide ranges of activities (particularly in buffered solutions). There are some experimental limitations, e.g., solvent purification and handling or use of a reference electrode without salt bridges, but there are important advantages. Solutes may be more soluble in such media, and redox properties of the species may be altered in comparison with aqueous solutions. Measurements of pH in nonaqueous solvents almost without exception use the glass electrode in combination with an appropriate reference electrode, frequently the silver/silver-chloride electrode. In general, the response of the glass electrode follows the Nernst expression in nonaqueous solvents and is an accurate representation of the changes in activity of hydronium ion. Unfortunately, few, if any, standard buffers are available to calibrate pH meters for nonaqueous measurements. Thus, nonaqueous pH measurements are only meaningful for monitoring the course of an acid-base titration or relative to some reference measurement made within the individual laboratory. Little, if any, confidence can be attached to absolute pH measurements in nonaqueous systems. The application of ion-selective electrodes in nonaqueous media has been limited, but the response for several cations and anions is usable, especially when 10–20% of water is added to pure nonaqueous solvent. Limited use of liquid-membrane electrodes in such media arises from the solubility of electrode-system components in organic media.

Second-class electrodes, that is those whose response is dependent on the change in concentration of an anion which gives an insoluble salt with the metal ion of the indicator electrode, provide a general means for monitoring the concentrations of anions. Some of those half reactions, which are well behaved electrochemically and provide means for the potentiometric monitoring of anion species, are summarized in Table I. This table also includes a tabulation of redox reactions that are useful to monitor the concentration of ligands that can complex metal ions. Consideration of these indicates one of the difficulties with absolute potentiometric measurements as a measure of metal-ion activity. If one is concerned purely with the actual activity of free metal ion, these measurements are meaningful. However, if the measurement is taken to represent the total metal-ion content of the solution, both as a free solvated ion and as its various complexes, then highly erroneous conclusions can be made. Thus, the ability to monitor the concentration of ligands should be recognized as a pitfall if one does not take account of this in the use of potential measurements to monitor metal-ion concentrations.

A recent and rapidly developing extension of potentiometry is in the area of membrane-type indicator electrodes. These include (1) specialized glass electrodes that respond to ions other than the hydronium ion, (2) ion-exchange membranes, and (3) single-inorganic-crystal membranes. Each year the selectivity and reliability are improved for this important class of electrode. In particular, the development of ion-exchange membranes that provide selective response for a number of anions has made new areas of analysis amenable to potentiometric measurements. This has been particularly important for the biomedical field where nondestructive, highly specific poteniometric measurements are desirable. Furthermore, the potentiometric method, because of its continuous nature, is particularly attractive to those concerned with *in vivo* monitoring of biological substances.

Because potentiometry (through the Nernst equation) gives a response that is proportional to the logarithm of the activity of the electroactive ion, the accuracy and precision are more limited than for many methods that give a direct proportional response. Thus, for a one-electron redox process, an order of magnitude change in activity gives a potential change of 59.1 mV (at room temperature), a 10% change in activity gives a change of 2.5 mV, and a 1% change in activity gives only a change of 0.25 mV.

Potentiometry has found extensive application over the past 50 years as a means to evaluate various thermodynamic parameters. Although this is not the major application of the technique today, it still provides one of the most convenient and reliable approaches to the evaluation of thermodynamic quantities. In particular, the activity coefficients of electroactive species can be evaluated directly through the use of the Nernst equation (for species that give a reversible electrochemical response). Thus, if an electrochemical system is used without a junction potential and with a reference electrode that has a well-established potential, then potentiometric measurement of the constituent species at a known concentration provides a direct measure of its activity. This provides a direct means to evaluate the activity coefficient (assuming the standard potential is known accurately for the constituent half reaction). If the standard half-reaction potential is not available, it must be evaluated under conditions

where the activity coefficient can be determined by the Debye-Huckel equation.

Another important application is the use of potentiometric measurements for the evaluation of thermodynamic equilibrium constants. In particular, the dissociation constants for weak acids and weak bases in a variety of solvents are evaluated conveniently with a pH electrode measuring system. The most precise approach is to perform an acid-base titration such that the titration curve can be recorded. Obviously, one could measure the pH of a known concentration of a weak acid and obtain a value of its hydronium-ion activity which would permit a direct evaluation of its dissociation constant. However, this would be a one-point evaluation and subject to greater errors than by titrating the acid halfway to the equivalence point. The latter approach uses a well-buffered region where the pH measurement represents the average of a large number of data points. Similar arguments can be made for the evaluation of solubility products and stability constants of complex ions.

In the use of potentiometry for the evaluation of stability constants for complex ions, the expressions can become extremely complicated if multi-equilibria are present. For a simple one-to-one complex, a direct potentiometric titration curve again provides the most satisfactory route to an accurate evaluation of the constant. The curve looks similar to that for an acid-base titration, and the appropriate point to pick is the half-equivalence point. If the complex is extremely stable, then the amount of free metal ion at this point on the titration curve (ligand titrated with metal ion) is sufficiently low that it can be disregarded. Assuming it is a stable complex, at the first half-equivalence point the concentration of complexed metal ion will be equivalent to that of the free ligand. The potential will give a direct measure of the free metal ion and allow the stability constant for the complex to be evaluated at the half-equivalence point, e.g.,

$$K_f = \frac{[\mathrm{Ag^I(en)^+}]}{\left[\mathrm{Ag^I(OH_2)_2^+}\right][\mathrm{en}]}. \tag{44}$$

Potentiometry also is a direct means to evaluate the standard potential for half reactions (E°) and has been applied for appropriate reversible systems. Such measurements require corrections for activity coefficients or extrapolation of the data to infinite dilution. Again, direct measurements in which equal molar concentrations of the oxidized or reduced species are introduced into the system provide a simple approach to such evaluations and are as precise as those obtained by less direct methods. However, E° values also can be extracted from potentiometric titration data. For example, in the titration of $\mathrm{Fe^{II}(OH_2)_6^{2+}}$ ion with $\mathrm{Ce^{IV}(OH_2)_n^{4+}}$ ion, the $\mathrm{Fe^{III}(OH_2)_6^{3+}}$ ion concentration equals the $\mathrm{Fe^{II}(OH_2)_6^{2+}}$ concentration at the half-equivalence point; the half-reaction potential, assuming activities are equal to concentrations, is given directly by the potential of the indicator electrode relative to the reference electrode. If the latter is a standard hydrogen electrode, the measured potential is equal to the E° for the $\mathrm{Fe^{III}(OH_2)_6^{3+}/Fe^{II}(OH_2)_6^{2+}}$ couple. The evaluation of the E° for a half reaction provides a direct measure of the free energy for the half reaction relative to the free energy for the reduction of hydronium ion to hydrogen gas. Likewise, a combination of any pair of E° values or of the free energy values permits the evaluation of the equilibrium constant and the standard free energy for a redox reaction.

From a practical standpoint it is often useful to have the observed potential in the medium of measurement for the condition of equal concentrations of the oxidized and reduced species of a half reaction. Such potentials are known as formal potentials, $E^{\circ\prime}$, rather than standard potentials and are not purely thermodynamic quantities. The term "formal potential" comes from the tradition of having the supporting electrolyte at a one formal concentration. However, other stated solution conditions also are included in many listings. Thus, the indicated potential is what one would expect at the half-equivalence point under actual titration conditions. In other words, activity corrections have not been made.

III. CONTROLLED-POTENTIAL ELECTROLYSIS AND VOLTAMMETRY

For chemists, the second important application of electrochemistry (beyond potentimetry) is the measurement of species-specific [e.g., iron(III) and iron(II)] concentrations. This is accomplished by an experiment whereby the electrolysis current for a specific species is independent of applied potential (within narrow limits) and controlled by mass transfer across a concentration gradient, such that it is directly proportional to concentration ($i = kC$). Although the contemporary methodology of choice is cyclic voltammetry, the foundation for all voltammetric techniques is polarography (discovered in 1922 by Professor Jaroslov Heyrovsky; awarded the Nobel Prize for Chemistry in 1959). Hence, a historical approach is used with a recognition that cyclic voltammetry will be the primary methodology for most chemists.

A. Principles and Fundamental Relations

1. Diffusion to a Planar Electrode

The basic approach in controlled-potential methods of electrochemistry is to control in some manner the potential of the working electrode while measuring the resultant current, usually as a function of time. When a potential

sufficient to electrolyze the electroactive species completely is applied to the electrode at ($t = 0$), the concentration at the electrode surface is reduced to zero and an electrode process occurs, for example,

$$ox + ne^- \rightleftharpoons red, \tag{45}$$

where *ox* and *red* represent an oxidized and reduced form of an electroactive species. Passage of current requires material to be transported to the electrode surface as well as away from it. Thus, relationships must be developed which involve the flux and diffusion of materials; this is appropriately accomplished by starting with Fick's second law of diffusion,

$$\frac{\partial C_{(x,t)}}{\partial t} = \frac{D\partial C_{(x,t)}}{\partial x^2}, \tag{46}$$

where D represents the diffusion coefficient, C represents the concentration of the electroactive species at a distance x from the electrode surface, and t represents the amount of time that the concentration gradient has existed. Through the use of Laplace transforms with initial and boundary conditions;

$$\text{for } t = 0 \text{ and } x \geq 0 \qquad C = C^b,$$

$$\text{for } t \geq 0 \text{ and } x \rightarrow 0 \qquad C \rightarrow C^b,$$

$$\text{for } t > 0 \text{ and } x = 0 \qquad C = 0.$$

Equation (46) can be solved to give a relationship for concentration in terms of parameters x and t,

$$C_{(x,t)} = C^b \operatorname{erf} \frac{x}{2D^{1/2}t^{1/2}}, \tag{47}$$

where C^b is the bulk concentration of the electroactive species.

By taking the derivative of Eq. (47) for the proper boundary condition, namely at the electrode surface ($x = 0$), the diffusion gradient at the electrode surface is expressed by the relation

$$\left(\frac{\partial C}{\partial x}\right)_{(0,t)} = \frac{C^b}{\pi^{1/2}D^{1/2}t^{1/2}}. \tag{48}$$

This flux of material crossing the electrode boundary can be converted to current by the expression

$$i = nFAD\frac{\partial C_{(0,t)}}{\partial x}, \tag{49}$$

where n is the number of electrons involved in the electrode reaction, F is the faraday, and A is the area of the electrode. When Eq. (48) is substituted into this relation a complete expression for the current that results from semi-infinite linear diffusion is obtained (the Cottrell equation for a planar electrode),

$$i = \frac{nFAC^bD^{1/2}}{\pi^{1/2}t^{1/2}}. \tag{50}$$

This relationship holds for any electrochemical process that involves semi-infinite linear diffusion and is the basis for a variety of electrochemical methods (e.g., polarography, voltammetry, and controlled-potential electrolysis). Equation (50) is the basic relationship used for solid-electrode voltammetry with a preset initial potential on a plateau region of the current-voltage curve. Its application requires that the electrode configuration be such that semi-infinite linear diffusion is the controlling condition for the mass-transfer process.

B. Voltammetry

1. Polarography

The most extensively studied form of voltammetry has been polarography (first described by Heyrovsky in 1922, with the quantitative relationships of current, potential, and time completed by the early 1930s with the assistance of associates such as Ilkovic). The potential-time dependence that is used for polarographic measurements is presented in Fig. 2 (solid line). The potential is scanned from E_1 to E_2 to obtain a current response that qualitatively and quantitatively characterizes the electroactive species present. The vast body of data from polarographic measurements can be adopted by other electroanalytical methods. Moreover, pulse polarographic methods and anodic stripping analysis, which are still used for determination of trace amounts of metal ions, are closely related to polarography. The unique characteristic of polarography is its use of a dropping mercury electrode, such that the electrode surface is continuously renewed in a well-defined and regulated manner to give reproducible effective electrode areas as a function of time. The diffusion current equation [Eq. (50)] can be extended to include a dropping mercury electrode by appropriate substitution for the area of the electrode. Thus, the volume of the drop for a dropping mercury electrode is given by the relationship

$$\mathrm{V} = \frac{4}{3}\pi r^3 = \frac{mt}{d}, \tag{51}$$

where r is the radius of the drop of mercury, m is the mass flow rate of mercury from the orifice of the capillary, t is the life of the drop, and d is the density of mercury under the experimental conditions. When this equation is solved for r and the latter is substituted into the equation for the area of a sphere, an expression for the area of the dropping mercury electrode drop as a function of the experimental parameters is obtained,

$$A = (4\pi)^{1/3}3^{2/3}m^{2/3}t^{2/3}d^{-2/3}. \tag{52}$$

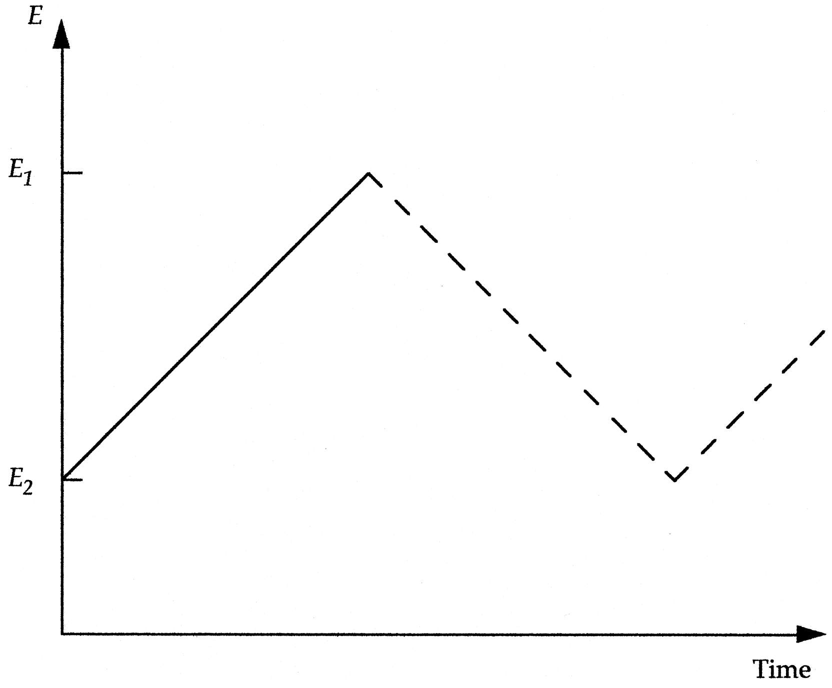


FIGURE 2 Potential-time profile that is used for polarography and linear-sweep voltammetry (solid line) and cyclic voltammetry (both solid and dashed lines).

This then can be substituted into Eq. (50) to give a calculated diffusion current for the dropping mercury electrode,

$$(i_t)_{\text{calc}} = 462nC^bD^{1/2}m^{2/3}t^{1/6}. \tag{53}$$

Actually, experimental results indicate that the constant in Eq. (53) is too small by a factor of $\sqrt{7/3}$, due to the growth of the mercury drop into the solution away from the capillary orifice. Thus, the correct diffusion current expression for a dropping mercury electrode is

$$i = 706nC^bD^{1/2}m^{2/3}t^{1/6}, \tag{54}$$

which gives the current at any time up to the lifetime of the drop. If the drop time, t_d, is substituted, the well-known Ilkovic equation results,

$$i_d = 706nC^bD^{1/2}m^{2/3}t_d^{1/6}, \tag{55}$$

where i_d is in A if D is in $cm^2\ s^{-1}$, C^b is in mol cm^{-3}, m in mg s^{-1}, and t is in s. Alternatively, i_d is in μA when C^b is in mM.

The polarographic current-potential wave that is illustrated by Fig. 3 conforms to the Nernst equation for reversible electrochemical processes. However, it is more convenient to express the concentrations at the electrode surface in terms of the current, i, and diffusion current, i_d. Because i_d is directly proportional to the concentration of the electroactive species in the bulk and i at any point on the curve is proportional to the amount of material produced by the electrolysis reaction, these quantities can be directly related to the concentration of the species at the electrode surface. For a generic reduction process [Eq. (45)], the potential of the electrode is given by the Nernst equation

$$E = E^\circ + \frac{RT}{nF}\ln\frac{C_{ox(0,t)}}{C_{red(0,t)}}. \tag{56}$$

Initially, the solution contains only *ox* (concentration, C^b_{ox}). When a potenital is applied and reductive current flows, the oxidized form of the electroactive species diffuses toward the electrode. From Eq. (63) it follows that

$$\begin{aligned} i &= 706n\left[C^b_{ox} - C_{ox(0,t)}\right]D^{1/2}_{ox}m^{2/3}t_d^{1/6} \\ &= i_d - 706nC_{ox(0,t)}D^{1/2}_{ox}m^{2/3}t_d^{1/6}. \end{aligned} \tag{57}$$

Hence,

$$C_{ox(0,t)} = \frac{i_d - i}{706nD^{1/2}_{ox}m^{2/3}t^{1/6}}. \tag{58}$$

After reduction of *ox*, its reduced form (*red*) diffuses either into the bulk of solution or into the mercury to form an amalgam. In either case,

$$i = 706n\left[C_{red(0,t)} - C^b_{red}\right]D^{1/2}_{red}m^{2/3}t_d^{1/6}. \tag{59}$$

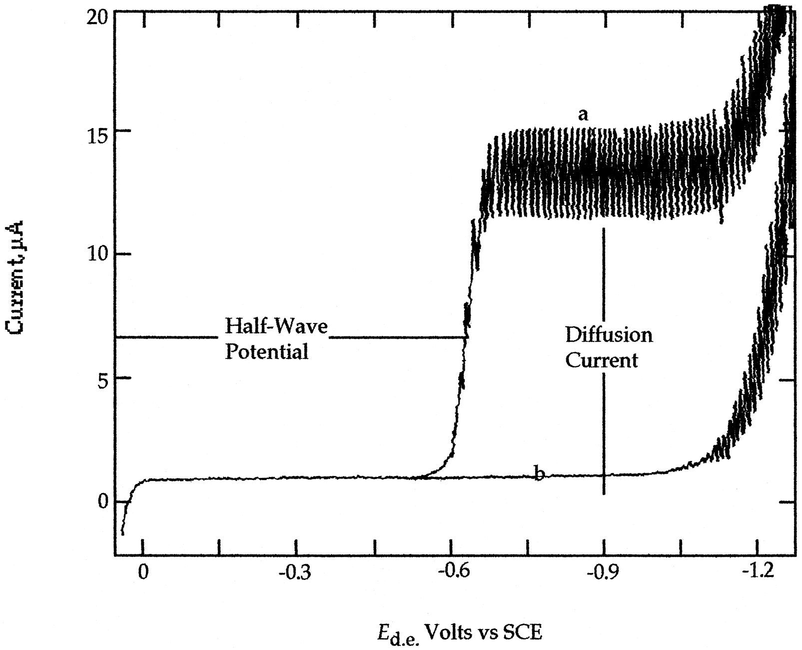


FIGURE 3 Polarograms for (a) 0.5 mM $Cd^{II}(OH_2)_6^{2+}$ in 1 M HCl and (b) 1 M HCl.

But the reduced form of the electroactive species was not present in the solution before electrolysis, therefore, $C^b_{red} = 0$ and

$$C_{red(0,t)} = \frac{i}{706nD^{1/2}_{red}m^{2/3}t^{1/6}}. \quad (60)$$

Substitution of Eqs. (58) and (60) into Eq. (56) gives

$$E = E^\circ + \frac{RT}{nF}\ln\left(\frac{D_{red}}{D_{ox}}\right)^{1/2} + \frac{RT}{nF}\ln\frac{i_d - i}{i}. \quad (61)$$

At the half-height of a polarographic wave ($i = i_d/2$), the corresponding potential is defined as the half-wave potential ($E_{1/2}$). Therefore, Eq. (61) takes the form

$$E = E_{1/2} + \frac{RT}{nF}\ln\frac{i_d - i}{i}. \quad (62)$$

For the reduction of a simple solvated metal ion to its amalgam, $E_{1/2}$ is given by

$$E_{1/2} = E^\circ + \frac{RT}{nF}\ln\frac{\gamma_{ion}D_a^{1/2}}{\gamma_a D_{ion}^{1/2}}, \quad (63)$$

where γ_{ion} is the activity coefficient for the ion and γ_a is the activity coefficient for the amalgamated species. The diffusion coefficients for the amalgam and ionic species also are a part of this expression. The standard reduction potential is for reduction of the ion to the amalgamated species. These expressions also hold for the reduction of an ion to a lower oxidation state, but require that the appropriate value be used.

2. Linear Sweep and Cyclic Voltammetry

The potential-time relation for voltammetric measurements is presented in Fig. 2. With linear-sweep voltammetry, the potential is linearly increased between potentials E_1 and E_2. Cyclic voltammentry is an extension of linear-sweep voltammetry with the voltage scan reversed after the current maximum (peak) of the reduction process has been passed. The voltage is scanned negatively beyond the peak and then reversed in a linear positive sweep. Such a technique provides even more information about the properties and characteristics of the electrochemical process and also gives insight into any complicating side processes such as pre- and post-electron-transfer reactions as well as kinetic considerations. Whereas in classical polarography the voltage-scan rate is about 1 V min^{-1}, linear-sweep voltammetry uses scan rates up to 100 V s^{-1} for conventional microelectrodes (and up to 10,000 V s^{-1} for ultra-microelectrodes; 10^{-6} m diameter).

Figure 4 illustrates the shape of a cyclic voltammogram with an electrode of fixed area. The voltammogram is characterized by a peak potential, E_p, at which the current reaches a maximum value, and by value of the peak current, i_p. When the reduction process is reversible the peak current is given by the relation

$$i_p = 0.4463nFA(Da)^{1/2}C^b \quad (64)$$

with

$$a = \frac{nF\nu}{RT} = \frac{n\nu}{0.026} \quad \text{at } 25^\circ\text{C}, \quad (65)$$

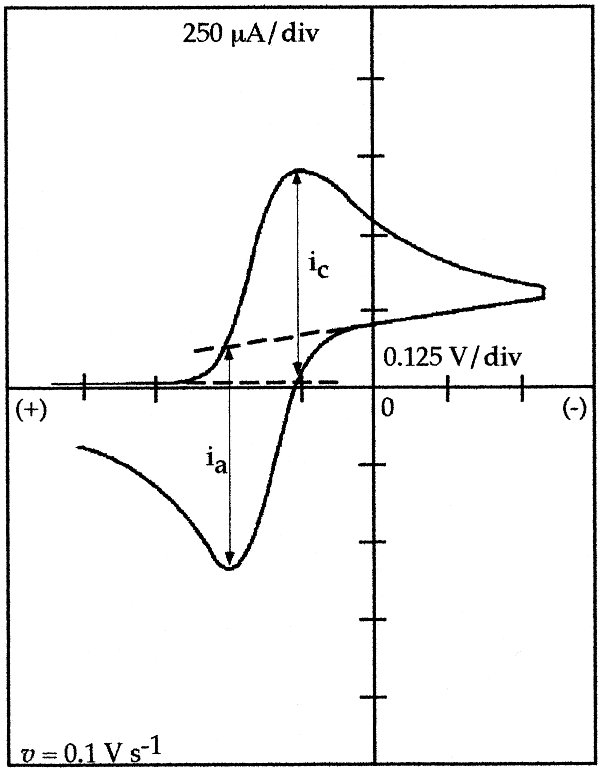


FIGURE 4 Linear voltage-sweep voltammogram with reversal of sweep direction to give a cyclic voltammogram. The initial sweep direction is to more negative potential.

where ν is the scan rate in volts per second. This relation results from the set of differential equations for Fick's second law of diffusion (with the appropriate initial and boundary conditions for *ox* and *red*). Thus, in terms of the adjustable parameters the peak current is given by the Randles-Sevcik equation

$$i_p = 2.69 \times 10^5 n^{3/2} A D^{1/2} C^b \nu^{1/2} \qquad \text{at } 25°\text{C}, \qquad (66)$$

where i_p is in A, A is in cm^2, D is in cm^2 s^{-1}, C^b is in mol cm^{-3}, and ν is in V s^{-1}.

Nicholson and Shain revolutionized the voltammetric experiment with their elegant development and demonstration of linear-sweep and cyclic voltammetry. In their approach, the current-potential curve is presented as

$$i = nFAC^b(\pi D a)^{1/2}\chi(at). \qquad (67)$$

From tabulations for the relation between $\pi^{1/2}\chi(at)$ and $n(E - E_{1/2})$ and converting the term $\pi^{1/2}\chi(at)$ to χ_{rev}, Eq. (67) takes the form

$$i = nFAC^b(Da)^{1/2}\chi_{rev}. \qquad (68)$$

For a given potential (E) the value of χ_{rev} is obtained from tabulations.

For a reversible process the peak potential can be related to the polarographic half-wave potential, $E_{1/2}$, by the expression

$$E_p = E_{1/2} - 1.11\frac{RT}{nF} = E_{1/2} - \frac{0.0285}{n} \qquad \text{at } 25°\text{C} \qquad (69)$$

Another useful parameter of the voltammetric curves is the half-peak potential, $E_{p/2}$, which is the potential at which the registered current reaches half its maximum value and is used to characterize a voltammogram. For a reversible process, $E_{1/2}$ is located halfway in between E_p and $E_{p/2}$.

The ratio of the peak current for the cathodic process relative to the peak current for the anodic process is equal to unity ($i_{p,c}/i_{p,a} = 1$) for a reversible electrode process. To measure the peak current for the anodic process the extrapolated baseline going from the foot of the cathodic wave to the extension of this cathodic current beyond the peak must be used as a reference, as illustrated by Fig. 4.

For the condition

$$|E_\lambda - E_{p/2}| \geq \frac{0.141}{n}, \qquad (70)$$

where E_λ is the extent of the voltage sweep, the difference in the peak potentials between the anodic and cathodic processes of the reversible reaction is given by the relationship

$$|\Delta E_p| = |E_{p,a} - E_{p,c}| = \frac{0.059}{n} \qquad \text{at } 25°\text{C}, \qquad (71)$$

which provides a rapid and convenient means to determine the number of electrons involved in the electrochemical reaction. For a reversible system, i_p is a linear function of $\sqrt{\nu}$, and E_p is independent of ν.

C. Controlled-Potential Bulk Electrolysis

Because of the extensive amount of data available from the polarographic and voltammetric literature, the optimum conditions for macroscopic electrolyses often are established. In particular, controlled-potential electrolysis at a mercury pool can be approached with predictable success on the basis of available polarographic information for the system of interest. An electrolysis can be accelerated by maximizing the electrode surface area and minimizing the thickness of the diffusion layer. However, the same electrode material must be used as in polarography. Thus, a conventional approach in controlled-potential electrolysis is the use of a mercury pool stirred as vigorously as possible with a magnetic stirring bar to minimize the concentration gradient. Under such conditions the decay of

the current as well as the decay of the concentration of the electroactive species is given by the relation

$$\frac{i_t}{i_{t=0}} = \frac{C_t}{C_{t=0}} = \exp\left[\left(\frac{-DA}{V\Delta x}\right)t\right], \qquad (72)$$

where V is the volume of the solution to be electrolyzed and Δx is the thickness of the concentration gradient. Thus, the current and concentration decay exponentially. Under idealized conditions, 90% of the electroactive species will be electrolyzed in approximately 20 min. Increases in the temperature as well as in the electrode area relative to the solution volume will accelerate the rate of electrolysis. The fundamentals of the controlled-potential bulk electrolysis are discussed in recent monographs (see Bibliography).

D. Applications of Controlled-Potential Methods

To date, the most extensive application of electrochemical methods with controlled potential has been in the area of qualitative and quantitative analysis. Because a number of monographs have more than adequately reviewed the literature and outlined the conditions for specific applications, this material is not covered here. In particular, inorganic applications of polarography and voltammetry have been discussed in great detail in the classic treatise by Kolthoff and Lingane.

An important specialized type of voltammetric system is a self-contained cell for the determination of $\cdot O_2\cdot$ in the gas or solution phases. This is the so-called Clark electrode, which consists of a platinum or gold electrode in the end of a support rod that is covered by an $\cdot O_2\cdot$-permeable membrane (polyethylene or Teflon) such that a thin film of electrolyte is contained between the electrode surface and the membrane. A concentric tube provides the support for the membrane and the means to contain an electrolyte solution in contact with a silver-silver chloride reference electrode. The Clark device has found extensive application to monitor $\cdot O_2\cdot$ partial pressure in blood, in the atmosphere, and in sewage plants. By appropriate adjustment of the applied potential, it gives a voltammetric current plateau that is directly proportional to the $\cdot O_2\cdot$ partial pressure. The membrane material prevents interference from electroactive ions as well as from surface-contaminating biological materials.

In addition to the analytical applications discussed above, controlled-potential methods are used for the evaluation of thermodynamic data and diffusion coefficients in both aqueous and nonaqueous solvents. Polarographic and voltammetric methods provide a convenient and straightforward means to the evaluation of the diffusion coefficients in a variety of media. The requirements are that the current be diffusion controlled, the number of electrons in the electrode reaction be known, and the concentration of the electroactive species and the area of electrodes be known. With these conditions satisfied, diffusion coefficients can be evaluated rapidly over a range of temperatures and solution conditions.

Voltammetric methods also provide a convenient approach to establish the thermodynamic reversibility of an electrode reaction and for the evaluation of the electron stoichiometry for the electrode reaction. As outlined in earlier sections, the standard electrode potential, the dissociation constants of weak acids and bases, solubility products, and the formation constants of complex ions can be evaluated from polarographic half-wave potentials, if the electrode process is reversible. Furthermore, studies of half-wave potentials as a function of ligand concentration provide the means to determine the formula of a metal complex.

The techniques of voltage sweep and cyclic voltammetry provide the analytical and physical chemical capabilities of classical voltammetry and, in addition, provide the means to perform these measurements much more rapidly for a broader range of conditions. Cyclic voltammetry is particularly useful for the rapid assessment of thermodynamic reversibility and for the evaluation of the stoichiometry for the electrode reaction.

IV. ELECTRON-TRANSFER PROCESSES

In electrochemical cells, electron transfer occurs within the electrode/solution interface, with electron removal (oxidation) at the anode and electron introduction (reduction) at the cathode. The current through the solution is carried by the ions of the electrolyte, and the voltage limits are those for electron removal *from* and electron insertion *into* the solvent/electrolyte {e.g., $H_2O/(H_3^+O_{aq})(ClO_{4aq}^-)$; $[Na^I(OH_2)_6^+](Cl_{aq}^-)$}

$$2\,H_2O \overset{-e^-}{\rightleftharpoons} [H_2O(H_2O\cdot)^+] \rightleftharpoons H_3^+O$$
$$+[HO\cdot](E^\circ)_{pH\,0},\ +2.72\ \text{V vs NHE}$$
$$(E^\circ)_{pH\,7},\ +2.31\ \text{V} \qquad (73)$$

$$HO^- \overset{-e^-}{\rightleftharpoons} [HO\cdot] \qquad (E^\circ)_{pH\,14},\ +1.89\ \text{V} \qquad (74)$$

$$Cl^- \overset{-e^-}{\rightleftharpoons} Cl\cdot \qquad E^\circ,\ +2.41\ \text{V}$$
$$E^{\circ\prime}_{MeCN},\ +2.24\ \text{V}. \qquad (75)$$

In the gas phase, electron removal from atoms is limited by their ionization potential (e.g., H·, 13.6 eV; K·, 4.3 eV; Na·, 5.1 eV; Cu·, 7.7 and 20.3 eV; Ag, 7.6 eV; Fe, 7.9, 16.2, and 30.7 eV). However, in the solution phase, electron removal (oxidation) from the solvent may be facilitated

by the presence of substrate atoms (rather than be from them).

For example, with water at pH 0 the process of Eq. (73) is shifted −4.82 V when hydrogen atoms are present,

$$[\mathrm{H}\cdot] + 2\,\mathrm{H_2O} \xrightleftharpoons{-e^-} \left[(\mathrm{H_2O})\mathrm{H_3^+O}\right] \rightleftharpoons \mathrm{H_3^+O} + \mathrm{H_2O} \qquad E^\circ, -2.10\,\mathrm{V}. \quad (76)$$

A. Oxidative Electrochemistry: Metals, Metal Complexes, Lewis Bases

1. Metals

The transformation of metal electrode surfaces via electrooxidation to their metallo-oxides, solvated ions, and metal complexes is fundamental to most anodic electrochemical processes (batteries, electrorefining, anodic stripping analysis, and reference electrodes). Although this is traditionally represented as the removal of one (or more) valence electron from a metal atom at the electrode surface to give a metal ion [e.g., Ag (s) $- e^- \rightarrow Ag^+$; E°, +0.80 V vs NHE], the gas-phase ionization potential [e.g., Ag · (g) $- e^- \rightarrow Ag^+$ (g); IP, 7.6 eV] is far greater than the observed oxidation potential. The difference is attributed to the solvation energy for the metal ion [e.g., $Ag^+ + n\mathrm{H_2O} \rightarrow Ag^+(aq)$; $-\Delta G(\mathrm{aq}) \approx 70$–100 kcal mol^{-1}]. However, such a sequential path would not obviate the 7.6-V energy barrier for the initial step and is in conflict with the observed thermodynamic reversibility for many metal/solvated-metal-ion redox couples.

All reactions, and particularly redox processes, occur via the easiest and lowest energy pathway that is available (mechanistically feasible) to the system. In the case of a metal electrode/electrolyte interface undergoing anodic transformations, the electrons can come from (**a**) surface metal atoms (energy limit; first ionization potential), (**b**) solvent molecules (energy limit; oxidation potential of solvent), (**c**) electrolyte anions (energy limit; oxidation potential of anions), and (**d**) base ligands (energy limit; oxidation potential of ligand). All metal electrodes are electrochemically transformed via path **b**, **c**, or **d** and never via path **a**. This general conclusion is illustrated for silver and copper electrodes in aqueous and acetonitrile (MeCN) solutions that contain inert electrolyte, chloride ion (Cl^-), or bipyridine (bpy).

In aqueous solutions at pH 0, the silver electrode facilitates oxidation of water

$$\mathrm{Ag\,(s)} + 2\,\mathrm{H_2O} \xrightarrow{-e^-} \mathrm{Ag^I(OH_2)_2^+} \qquad E^\circ,\ +0.80\,\mathrm{V\ vs\ NHE}. \quad (77)$$

The gas-phase ionization potential for a silver atom is 7.6 eV. In contrast, water is oxidized (gives up an electron) at much lower potentials [Eq. (73); at pH 0, +2.72 V vs NHE]. At a silver electrode, water oxidation is facilitated via formation of an Ag^I–OH_2^+ bond; the shift in oxidation potential from +2.72 to +0.80 V is a measure of the bond-formation energy ($-\Delta G_{BF}$),

$$-\Delta G_{\mathrm{BF}} = (+2.72 - 0.80)\,23.1 = 44.4\,\mathrm{kcal\ mol^{-1}}. \quad (78)$$

At pH 14 the anodic process is the oxidation of HO^-,

$$\mathrm{HO^-} \xrightarrow{e^-} \mathrm{HO}\cdot \qquad E^\circ, +1.89\,\mathrm{V} \qquad (E^\circ)_{\mathrm{pH\,7}}, +2.30\ \mathrm{V}, \quad (79)$$

which is facilitated via formation of an Ag^I–OH bond [$-\Delta G_{BF} = (1.89 - 0.34)\,23.1 = 35.8$ kcal mol^{-1}]

$$\mathrm{Ag\,(s)} + \mathrm{HO^-} \xrightarrow{-e^-} \mathrm{Ag^I{-}OH\,(s)} \qquad E^\circ, +0.34\ \mathrm{V} \qquad (E^\circ)_{\mathrm{pH\,7}}, +0.75\,\mathrm{V}. \quad (80)$$

The data of Eqs. (77) and (79) can be combined to give a value for the solubility product (K_{sp}) for Ag–OH (s),

$$\mathrm{Ag^I(OH_2)_2^+} + \mathrm{HO^-} \longrightarrow \mathrm{Ag^I{-}OH\,(s)} + 2\,\mathrm{H_2O}$$
$$\left[\mathrm{Ag^I(OH_2)_2^+}\right][\mathrm{HO^-}] = K_{\mathrm{sp}};$$
$$\log K_{\mathrm{sp}} = (0.34 - 0.80)/0.059 = -7.8. \quad (81)$$

In the presence of chloride ion, metal electrodes facilitate its oxidation [Eq. (75); E°, +2.41 V vs NHE] via formation of metal-chlorine covalent bonds, e.g.,

$$\mathrm{Ag\,(s)} + \mathrm{Cl^-} \xrightarrow{-e^-} \mathrm{Ag{-}Cl\,(s)} \qquad E^\circ, +0.22\,\mathrm{V}. \quad (82)$$

Hence, the differential bond-formation energy [$\Delta(-\Delta G_{BF})$] [Ag–Cl bond energy, minus the energy required to break an Ag–Ag bond at the Ag (s) surface)] is given by the difference in oxidation potentials [Eqs. (75) and (82)],

$$\Delta(-\Delta G_{\mathrm{BF}}) = (2.41 - 0.22)\ 23.1 = 50.6\,\mathrm{kcal\ mol^{-1}}. \quad (83)$$

Because the escape energy for an Ag· atom from Ag (s) is 68 kcal mol^{-1}, a reasonable approximation for the breakage of a single bond is 22.6 kcal mol^{-1} [(1/3) 68]. When combined with Eq. (83), this gives a reasonable value for $-\Delta G_{BF}$,

$$\mathrm{Ag}\cdot + \mathrm{Cl}\cdot \longrightarrow \mathrm{Ag{-}Cl} \qquad -\Delta G_{\mathrm{BF}} = 50.6 + 22.6 = 73.2\,\mathrm{kcal\ mol^{-1}}. \quad (84)$$

The literature value for the dissociative bond energy (ΔH_{DBE}) of Ag–Cl (g) is 81.6 kcal mol^{-1}, which is equivalent to an estimated $-\Delta G_{BF}$ value of 73.8 kcal mol^{-1}

$[-\Delta G_{BF} = \Delta H_{DBE} - T\Delta S = 81.6 - 7.8\,(est) = 73.8$ kcal mol^{-1}]. Thus, the proposition that metal electrode oxidations are solvent or ligand centered with potentials that reflect the metal-solvent/ligand bond-formation free energies ($-\Delta G_{BF}$) is supported by independent bond-energy data. The data of Eqs. (77) and (82) provide a measure of the solubility product for AgCl (s).

$$Ag^{I}(OH_2)_2^+ + Cl^- \longrightarrow Ag^{I}{-}Cl\,(s) + 2\,H_2O$$

$$[Ag^{I}(OH_2)_2^+][Cl^-] = K_{sp};$$

$$\log K_{sp} = (0.22 - 0.80)/0.059 = -9.8. \quad (85)$$

Another important example is the oxidation of Cl^- at a mercury electrode [Hg_2 (l)] to form calomel [mercurous chloride, Hg_2Cl_2 (s), Cl–Hg^{II}–Hg^{II}–Cl (s)].

$$Hg_2\,(l) + Cl^- \xrightarrow{-e^-} [Cl{-}Hg{-}Hg\cdot]$$

$$\xrightarrow[Cl^-]{-e^-} Cl{-}Hg{-}Hg{-}Cl\,(s) \qquad E^\circ, +0.27\,V. \quad (86)$$

The potential shift for the $Cl^-/Cl\cdot$ couple from +2.41 [Eq. (75)] to +0.27 V in the presence of Hg_2 (l) is a measure of the [Cl–HgHg] bond energy [$-\Delta G_{BF} = (2.41 - 0.27)\,23.1 = 49.4$ kcal mol^{-1}].

Similar metal-facilitated oxidations of H_2O and of Cl^- occur for all metal electrodes. The respective potentials for the oxidation of each at a copper electrode are

$$Cu\,(s) + 2\,H_2O \overset{-e^-}{\rightleftharpoons} (H_2O)Cu^{I}{-}OH_2^+ \quad (E^\circ)_{pH\,0}, +0.52\,V$$

$$[-\Delta G_{BF}, 46\,\text{kcal mol}^{-1}] \quad (87)$$

and

$$Cu\,(s) + Cl^- \overset{-e^-}{\rightleftharpoons} Cu^{I}{-}Cl \qquad E^\circ, +0.14\,V$$

$$[-\Delta G_{BF}, 52\,\text{kcal mol}^{-1}]. \quad (88)$$

Additional redox data for oxidations of H_2O/HO^- at Cu, Ag, and Au electrodes in aqueous and acetonitrile (MeCN) solutions are summarized in Table I. At pH 0 with an iron electrode the water oxidation of Eq. (73) is shifted by −3.12 V,

$$Fe\,(s) + 6\,H_2O \overset{-2e^-}{\rightleftharpoons} [(H_2O)_4Fe^{II}(OH_2)_2]^{2+}$$

$$(E^\circ)_{pH\,0}, -0.40\,V, \quad (89)$$

which indicates that the $[H_2O(H_2O\cdot)^+]$ species is stabilized by a strong $[(H_2O)_4(H_2O)^+]Fe^{II}$–$OH_2^+$ covalent bond ($-\Delta G_{BF}$, ~71 kcal mol^{-1}).

2. Metal Complexes

In an analogous fashion, the removal of an electron (oxidation) from water via Eq. (73) is aided by the presence of transition-metal ions [e.g., $Cu^{I}(OH_2)_6^+$, $Fe^{II}(OH_2)_6^{2+}$, and $Ce^{III}(OH_2)_6^{3+}$, each with one, two, and three M–OH_2^+ covalent bonds, respectively],

$$Cu^{I}(OH_2)_6^+ + H_2O \xrightarrow[(E^\circ)_{pH\,0}, +0.16\,V]{-e^-} [(H_2O)_5^+]Cu^{II}$$

$$-OH_2^+(OH_2) \longrightarrow (H_2O)_5^+Cu^{II}{-}OH + H_3^+O$$

$$[-\Delta G_{BF}, 59\,\text{kcal mol}^{-1}] \quad (90)$$

$$Fe^{II}(OH_2)_6^{2+} + H_2O \xrightarrow[(E^\circ)_{pH\,0}, +0.77\,V]{-e^-} [(H_2O)_5^{2+}]Fe^{III}$$

$$-OH_2^+(OH_2) \longrightarrow (H_2O)_5^{2+}Fe^{III}{-}OH + H_3^+O$$

$$[-\Delta G_{BF}, 45\,\text{kcal mol}^{-1}] \quad (91)$$

$$Ce^{III}(OH_2)_6^{3+} + H_2O \xrightarrow[(E^\circ)_{pH\,0}, +1.60\,V]{-e^-} [(H_2O)_5^{3+}Ce^{IV}$$

$$-OH_2^+(OH_2)] \longrightarrow (H_2O)_5^{3+}Ce^{IV}{-}OH + H_3^+O$$

$$[-\Delta G_{BF}, 25\,\text{kcal mol}^{-1}]. \quad (92)$$

In none of these examples has the potential for removal of an electron approached the ionization potentials of the metals. Although traditional treatments attribute the potentials of Eqs. (77), (87), and (89)–(92) to the removal of electrons from the metals, coupled with large ionic solvation energies, this requires a pathway with the ionization potential as a kinetic barrier. Furthermore, the spontaneous reaction of iron with acidified water is driven by the formation of Fe–OH_2^+ and H–H covalent bonds that facilitate hydrogen-atom transfer from water (rather than electron transfer from iron),

$$Fe\,(s) + 2\,H_3^+O \xrightarrow{H_2O} (H_2O)_4Fe^{II}(OH_2)_2^{2+} + H_2 \quad (93)$$

{Note: To ionize a gas-phase iron atom ($Fe \xrightarrow{-3e^-} Fe^{3+}$) requires 54.8 eV (1266 kcal mol^{-1}); in turn, this species reacts upon dissolution into liquid water {$Fe^{3+}(g) + 7H_2O(l) \rightarrow [(H_2O)_5^{2+}]Fe^{III}$–OH + H_3^+O, $-\Delta H \approx 1000$ kcal mol^{-1} (1266 − 266)}; the net energy change often is ascribed as the solvation energy for Fe^{3+} (g) (heat of hydration)}.

Within an aprotic solvent (e.g., acetonitrile, MeCN) oxidation of metals and metal complexes also is ligand centered with the potential determined by the oxidation potential of the ligand and the metal-ligand covalent bond-formation free energy ($-\Delta G_{BF}$). For example, the free bpy ligand in acetonitrile is oxidized near the solvent limit at a glassy-carbon electrode (GC) (Fig. 5),

$$(bpy) \xrightarrow[MeCN]{} bpy^{+\cdot} \qquad E^{\circ\prime}, +2.32\,V \text{ vs NHE}, \quad (94)$$

but at a copper electrode the oxidation occurs at a negative potential,

$$Cu\,(s) + 2\,bpy \xrightarrow{-e^-} Cu^{I}(bpy)_2^+ \qquad E^{\circ\prime}, -0.16\,V. \quad (95)$$

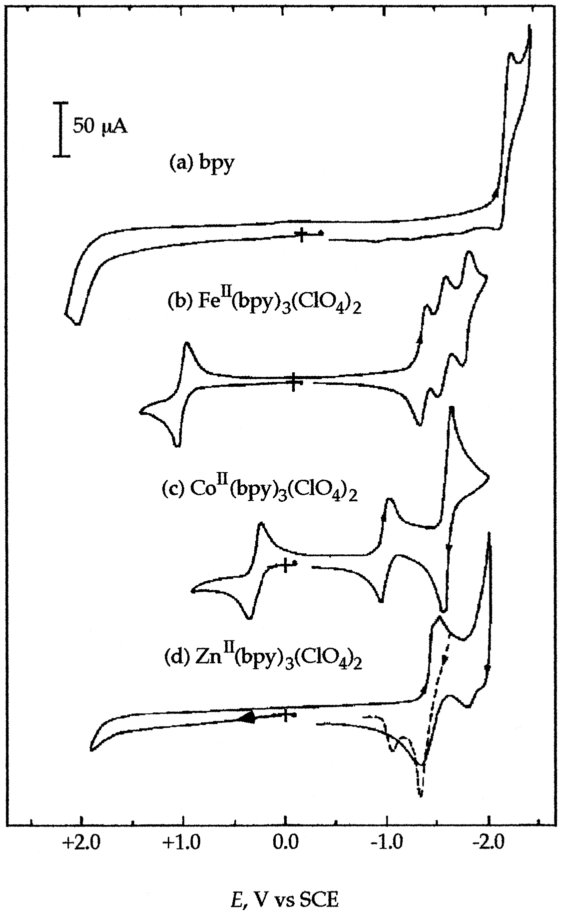


FIGURE 5 Cyclic voltammograms of 3 mM solution in MeCN (0.1 M tetraethylammonium perchlorate [TEAP]): (a) bpy; (b) $Fe^{II}(bpy)_3^{2+}$; (c) $Co^{II}(bpy)_3^{2+}$; (d) $Zn^{II}(bpy)_3^{2+}$. Conditions: scan rate, 0.1 V s^{-1}; 25°C; glassy-carbon working electrode (0.09 cm^2); SCE vs NHE, +0.244 V.

Figure 5 illustrates that the oxidation of the $Fe^{II}(bpy)_3^{2+}$ complex is reversible and ligand centered,

$$\underset{d^6sp}{Fe^{II}(bpy)_3^{2+}} \xrightarrow[\text{MeCN}]{-e^-} \underset{d^5sp^2}{Fe^{III}(bpy)_3^{3+}} \qquad E^{\circ\prime}, +1.30\ \text{V}. \qquad (96)$$

(Noteworthy are the three reversible one-electron reductions for this complex.) The electron that is removed from the $Fe^{II}(bpy)_3^{2+}$ complex comes from the ligands to give $bpy^{+}\cdot$, which couples with one of the four unpaired electrons of the iron(II) center (d^6sp) to give a third covalent bond [$Fe^{III}(bpy)_3^{3+}$, d^5sp^2; $S = 5/2$]. The difference in oxidation potentials for $Fe^{II}(bpy)_3^{2+}$ and free bpy [Eq. (94)] is a measure of the Fe^{III}–bpy^{+} bond energy [$-\Delta G_{BF} = (2.32 - 1.30)\ 23.1 = 23.6$ kcal mol^{-1}]. The potential that would be required to remove an electron from the d^6sp manifold of the iron(II) center of $Fe^{II}(OH_2)_6^{2+}$ or $Fe^{II}(bpy)_3^{2+}$ would be greater than the first ionization potential of atomic iron (7.9 eV).

TABLE II Oxidation Potentials for Ligands (L) and Their ML_3 Complexes with Zn(II), Mn(II), Fe(II), and Co(II) in MeCN (0.1 M Tetraethylammonium Perchlorate)

Ligand	$E^b_{1/2}$, V vs NHE^c				
(L or L^-)^a	$L/L^+\cdot$ ($L^-/L\cdot$)	$Zn^{II}L_3$	$Mn^{II}L_3$	$Fe^{II}L_3$	$Co^{II}L_3$
H_2O	+2.8	+2.8	+2.8	+1.84	+2.8
bpy	+2.32	>+2.5	+1.55	+1.30	+0.58
PA^-	+1.50	+1.54	+0.60	+0.20	+0.04
$acac^-$	+0.55	+0.58	+0.18	−0.42	−0.35
8-Q^-	+0.21	+0.22	−0.06	−0.41	−0.57

[a] Key: bpy, 2,2′-bipyridine; PA^-, picolinate (2-carboxylate pyridine); $acac^-$, acetylacetonate; 8-Q^-, 8-quinolinate.

[b] $E_{1/2}$ taken as $(E_{p,a} + E_{p,c})/2$ for reversible couples of $Mn^{II}L_3$ and $Fe^{II}L_3$ complexes; as $E_{p,a/2} + 0.03$ V for L (or L^-) and $Zn^{II}L_3$; and as $E_{p,c/2} - 0.03$ V for the $Co^{II}L_3$ complex that exhibits separated redox couples.

[c] SCE vs NHE; +0.242 V.

Table II summarizes the oxidation potentials for ligands (L) and their $M^{II}L_3$ complexes with zinc(II), manganese(II), iron(II), and cobalt(II). The difference in the potentials for the free and complexed ligands is a measure of the metal(III)-ligand covalent-bond-formation energies ($-\Delta G_{BF}$); these are summarized in Table III. All of the data are consistent with ligand-centered redox processes.

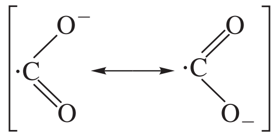
B. Reductive Electrochemistry

1. Free Electron (e^-); Ultimate Base and One-Electron Reductant in Any Matrix

The free electron interacts with all atoms and molecules that have finite electron affinities to produce anions and, thus, is unstable in all but the most inert liquids. Electrochemistry attests to this general axiom and provides a convenient means to evaluate the energetics for the addition of an electron to solvent molecules and to species within a solution, e.g.,

$$e^- + H_2O(aq) \longrightarrow [H\cdot] + HO^-(aq) \qquad E^\circ, -2.93\ \text{V vs NHE} \qquad (97)$$

$$e^- + CO_2\,(MeCN) \longrightarrow$$



$$E^\circ, -1.86\ \text{V vs NHE} \qquad (98)$$

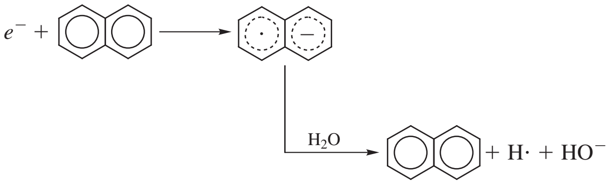
$$e^- + Me_2\ddot{S}(O) \longrightarrow H_2C{=}\ddot{S}(O^-)Me + [H\cdot] \qquad (99)$$

TABLE III Apparent Metal-Ligand Covalent-Bond-Formation Free Energies ($-\Delta G_{BF}$) for Several Manganese, Iron, and Cobalt Complexes

Complex	$-\Delta G_{BF}$, kcal mol^{-1} [a]
A. Manganese	
$(8Q)_2Mn^{III}$–8Q	6
$(acac)_2Mn^{III}$–acac	9
$(PA)_2Mn^{III}$–PA	22
$[(bpy)_2Mn^{III}–bpy]^{3+}$	>23[b]
B. Iron	
$(8Q)_2Fe^{III}$–8Q	15
$(acac)_2Fe^{III}$–acac	23
$(PA)_2Fe^{III}$–PA	31
$[(bpy)_2Fe^{III}–bpy]^{3+}$	>29[b]
$[(Ph_3PO)_3Fe^{III}–OPPh_3]^{3+}$	>30[b]
$[(MeCN)_4Fe^{III}–OH_2]^{3+}$	23
C. Cobalt	
$(8Q)_2Co^{III}$–8Q	16
$(acac)_2Co^{III}$–acac	21
$(PA)_2Co^{III}$–PA	35
$[(bpy)_2Co^{III}–bpy]^{3+}$	>46[b]

[a] $(-\Delta G_{BF}) = [E_{1/2[ZnL_3^-/ZnL_2(L\cdot)]} - E_{1/2(ML_3^-/M(\cdot L)L_2)}] \times 23.1$ kcal mol^{-1}.

[b] $(-\Delta G_{BF}) = [E_{p,a\,(ZnL/ZnL^{+}\cdot)} - E_{p,a(ML/M-L^{+})}] \times 23.1$ kcal mol^{-1}; $L = (bpy)_3$ or $(Ph_3PO)_4$.



(100)

Hence, reductive electrochemistry converts electrons (e^-) via the solution matrix at the interface to atoms and anions. The solution outside the inner double layer *never* is exposed to an electron. Some examples of such *inner-double-layer electron transfer* include

$$H_2O + e^- \longrightarrow [H\cdot] + HO^- \qquad E^\circ, -2.93 \text{ V vs NHE} \quad (101)$$

$$H_3^+O + e^- \longrightarrow [H\cdot] + H_2O \qquad E^\circ, -2.10\text{ V} \quad (102)$$

$$(H_2O)_5^2 + Fe^{III}\text{–}OH + H_3^+O + e^- \xrightarrow[H_2O]{pH\,1} Fe^{II}(OH_2)_6^{2+} + H_2O \qquad E^{\circ\prime}, +0.71\text{ V} \quad (103)$$

$$[HO\cdot] + H_3^+O + e^- \xrightarrow[H_2O]{pH\,1} 2\,H_2O \qquad E^{\circ\prime}, +2.66\text{ V} \quad (104)$$

$$-\Delta G_{BF}\left[(H_2O)_5^{2+}Fe^{III}\text{–}OH\right] = [2.66 - 0.71] \times 23.1 = 45 \text{ kcal mol}^{-1} \quad (105)$$

The electrochemical reduction of permanganic acid $[HOMn^{VII}(O)_3]$, which is traditionally represented as a metal-centered electron transfer to change Mn^{7+} to Mn^{6+}, is another example of a ligand-centered process,

$$(O)_3Mn^{VII}\text{–}OH + H_3^+O + e^- \longrightarrow Mn^{VI}(O)_3 + 2\,H_2O$$
$$-\Delta G_{BF} = 28 \text{ kcal mol}^{-1} \text{ (pH 1)}. \quad E^{\circ\prime}, +1.45\text{ V} \quad (106)$$

Comparison of this with the reduction of free hydroxyl radical (HO·) [Eq. (104)] provides a measure of the $HO\text{–}Mn^{VII}(O)_3$ bond energy [$-\Delta G_{BF} = (2.66 - 1.45)\,23.1 = 28$ kcal mol^{-1}]. The other strong oxidants [$(HO)_2Cr_2^{VI}(O)_5$ and $HOCe^{IV}(OH_2)_5^{3+}$] that are used for aqueous redox titrations are reduced by a similar path,

$$HO\text{–}Cr^{VI}(O)_2OCr^{VI}(O)_2OH + H_3^+O + e^- \xrightarrow[+1.30\text{ V}]{} (O)_2Cr^{V}OCr^{VI}(O)_2OH + 2\,H_2O$$
$$-\Delta G_{BF}, 31 \text{ kcal mol}^{-1} \text{ (pH 1)} \quad (107)$$

$$HO\text{–}Ce^{IV}(OH_2)_5^{3+} + H_3^+O + e^- \xrightarrow[+1.66\text{ V}]{} Ce^{III}(OH_2)_6^{3+} + H_2O$$
$$-\Delta G_{BF}, 23 \text{ kcal mol}^{-1} \text{ (pH 1)}. \quad (108)$$

An important point in these electron-transfer reductions is that the primary electron acceptor is the hydronium ion (H_3^+O), which is transformed to a hydrogen atom (H·) that reacts with HO· (either free or bound via a covalent bond to the metal center). Thus, in the reactions of Eqs. (103), (104), and (106)–(108), the oxidant in each is the hydronium ion (H_3^+O) and the reduction potential is determined by the H–OH bond energy ($-\Delta G_{BF}$) of the product H_2O, minus the metal–OH bond energy [Eqs. (106)–(108)].

Under alkaline conditions $Mn^{VII}O_4^-$ is reduced via direct electron addition to one of the bound oxygen atoms,

$$^-OMn^{VII}(O)_3 + e^- \longrightarrow {}^-OMn^{VI}(O)_2O^- \qquad E^\circ, +0.55\text{ V vs NHE}. \quad (109)$$

The extent of the stabilization of the oxygen atom in $Mn^{VII}O_4^-$ is indicated by the reduction potential for a free ·O· atom,

$$[\cdot O\cdot] + e^- \longrightarrow \cdot O^- \qquad (E^\circ)_{pH\,14}, +1.43\text{ V}. \quad (110)$$

In summary, the electron-transfer reactions for metals and metal complexes are ligand centered (or solvent centered). In each case the potential for the oxidation of free ligand is decreased in the presence of metal or reduced-metal complex by an amount that is proportional to the metal-ligand bond energy ($-\Delta G_{BF}$).

2. Electro-Induced Hydrogenation

Reductive electron transfer in an electrochemical cell occurs by insertion of an electron from the electrode (cathode) into the solution matrix within the double layer of the electrode/solution interface {e.g., $H_2O/(H_3^+O)(ClO_4^-)$; $[Na^I(OH_2)_6^+](ClO_4^-)$}.

$$H_3^+O \overset{e^-}{\rightleftarrows} [H\cdot] + H_2O \quad (E^\circ)_{pH\,0}, -2.10\text{ V vs NHE} \quad (E^\circ)_{pH\,7}, -2.51\text{ V}. \qquad (111)$$

The reductive processes of Eq. (111) may be facilitated by the presence of substrates to stabilize the H-atom product. For example, in pH 0 water, the reduction process is shifted by +4.82 V when hydroxyl radicals (HO·) are present,

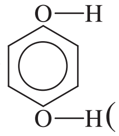
$$H_3^+O + [HO\cdot] \overset{e^-}{\rightleftarrows} HO\text{–}H + H_2O\ E^\circ, +2.72\text{ V} \quad (-\Delta G_{BF}, 111\text{ kcal mol}^{-1}) \qquad (112)$$

and by +2.10 V at a platinum electrode,

$$H_3^+O + Pt \overset{e^-}{\rightleftarrows} Pt\text{–}H + H_2O \quad E^\circ, 0.00\text{ V} \quad (-\Delta G_{BF}, 48\text{ kcal mol}^{-1}) \qquad (113)$$

In the presence of benzoquinone (Q), the shift is +2.80 V,

$$2H_3{}^+O + Q \overset{2e^-}{\rightleftarrows}$$



$$(E^\circ)_{pH\,0}, +0.70\text{ V} \quad (-\Delta G_{BF}, 65\text{ kcal mol}^{-1}). \qquad (114)$$

V. ELECTROCHEMICAL CHARACTERIZATION OF MOLECULES

An increasing number of chemists use electrochemistry as a characterization technique in a fashion analogous to their use of infrared, UV-visible, NMR, and ESR spectroscopy. Some of the chemical questions that are amenable to treatment by electrochemistry include (1) the standard potentials (E°) of the compound's oxidation-reduction reactions, (2) evaluation of the solution thermodynamics of the compound, (3) determination of the electron stoichiometry of the compound's oxidation-reduction reactions, (4) preparation and study of unstable intermediates, (5) evaluation of the valence of the metal in new compounds, (6) determination of the formulas and stability constants of metal complexes, (7) evaluation of M–X, H–X, and O–Y covalent-bond-formation energies ($-\Delta G_{BF}$), and (8) studies of the effects of solvent, supporting electrolyte, and solution acidity upon oxidation-reduction reactions.

Often, the first step in the electrochemical characterization of a compound is to ascertain its oxidation-reduction reversibility. Cyclic voltammetry usually is the most convenient and reliable technique for this and related qualitative characterizations of a new system. The discussion in earlier sections outlines the specific procedures and relationships. The next step in the characterization usually is the determination of the electron stoichiometry of the oxidation-reduction steps of the compound. Controlled-potential coulometry provides a rigorously quantitative means to such evaluations.

The electrochemical characterization of the oxidation/reduction chemistry of metals, metal ions, and metal complexes has been discussed in Section III. In the following paragraphs, the use of electrochemistry for molecular characterization is further illustrated.

A. Hydronium Ions, Brønsted Acids, and Molecular Hydrogen

The most fundamental redox process in electrochemistry is the reductive transformation of hydronium ion $[H_3^+O(aq)]$ at a platinum electrode to molecular hydrogen $[H_2\,(g)]$,

$$2\,H_3^+O(aq) + 2e^- \overset{Pt}{\rightleftarrows} H_2\,(g) \quad E^\circ, 0.0000\text{ V vs NHE}. \qquad (115)$$

When properly engineered and with $[H_3^+O(aq)]$ at unit activity and P_{H_2} at unit fugacity, this electrode system is the thermodynamic reference standard for measurements of electrochemical potentials and is referred to as the *Normal Hydrogen Electrode (NHE)*, which is alternatively called the *Standard Hydrogen Electrode (SHE)*,

$$E = E^\circ_{NHE} + 2.30\,RT/2F \log\left\{\left[H_3^+O(aq)\right]^2 / P_{H_2}\right\}$$
$$(E^\circ_{NHE} \equiv 0.0000\text{ V at all temperatures, 0–100°C})$$
$$[2.30RT/2F = 0.05915/2 \text{ at } 25°C]. \qquad (116)$$

The latter equation for the NHE also is the defining basis for the potentiometric measurement of hydronium-ion activity $[H_3^+O]$ and molecular hydrogen fugacity (P_{H_2}),

$$pH_a \equiv -\log\left[H_3^+O\right] = [(E^\circ_{NHE} - E_{ind})/0.059] - \log\left(P_{H_2}\right) = -E_{ind}/0.059\,(E^\circ_{NHE} = 0.000\text{ V}; P_{H_2} = 1.00\text{ atm}) \qquad (117)$$

$$\log\left(P_{H_2}\right) = [(E^\circ_{NHE} - E_{ind})\,2/0.059] + \log\left[H_3^+O\right]^2 = -E_{ind}\,(2/0.059)\left\{E^\circ_{NHE} = 0.000\text{ V}; \left[H_3^+O\right] = 1.000\right\}. \qquad (118)$$

Electrochemical measurement of pH_a via Eq. (116) senses hydronium-ion activity rather than its concentration. Hence, electrochemical evaluations of dissociation constants (K_{HA}) yield thermodynamic quantities

$$HA + H_2O \rightleftharpoons H_3^+O + A^- \quad K_{HA}$$
$$K_{HA} = [H_3^+O][A^-]/[HA]. \qquad (119)$$

1. Hydronium-Ion (H_3^+O) Reduction

Although the NHE is fundamental to electrochemistry, it does not represent the primary electron-transfer step for hydronium-ion reduction at an inert (glassy-carbon) electrode,

$$H_3^+O(aq) + e^- \xrightleftharpoons{GC} H\cdot(aq) \qquad E^\circ, -2.10 \text{ V vs NHE}, \qquad (120)$$

where GC means a glassy-carbon electrode. The −2.10-V difference in standard potential (E°) between the latter and that for the NHE [Eq. (115)] is due to the platinum electrode, which stabilizes the hydrogen atom (H·) via formation of a Pt–H covalent bond,

$$H_3^+O(aq) + Pt(s) + e^- \rightarrow Pt\text{–}H(s) \qquad E^\circ, 0.000 \text{ V}$$
$$\Delta(-\Delta G_{BF}) = [0.00 - (-2.10)]\ 96.48 \text{ kJ mol}^{-1} (\text{eV})^{-1}$$
$$= 203 \text{ kJ mol}^{-1} = 48 \text{ kcal mol}^{-1}, \qquad (121)$$

where $-\Delta G_{BF}$ is the free energy of bond formation.

In this system the platinum electrode surface does not consist of free platinum atoms, but must undergo homolytic Pt–Pt bond breakage [ΔH_{DBE}, 24 kcal mol^{-1} per Pt· (est)] before a Pt–H bond is formed. Thus, the Pt–H bond-formation energy ($-\Delta G_{BF}$) is estimated to be 72 kcal mol^{-1} on the basis of the electrochemical data [for the gas-phase Pt–H molecule, $\Delta H_{DBE} \leq 80$ kcal mol^{-1} or ($-\Delta G_{BF} \approx 72$ kcal mol^{-1})]. Determination of the reduction potential for H_3^+O(aq) at other metal electrodes (M) provides a convenient means to estimate M–H bond energies [$\Delta(-\Delta G_{BF})$].

2. Brønsted-Acid (HA) Reduction and Evaluation of $pK_{a(sol)}$

Brønsted acids (HA) undergo dissociation in any solvent to yield the solvated aqua-hydronium ion [$(sol)_nH_3^+O(aq)$] {from residual H_2O; or [$(sol)_nHsol^+$] for basic solvents}, which further dissociates to H_3^+O(aq) and the solvated conjugate base [$A^-(sol)_n$]

$$HA(sol) + H_2O \rightleftharpoons H_3^+O(aq) + A^-(sol) \quad (K_a)_{sol} \qquad (122)$$

and

$$H_3^+O(aq) = (K_a)_{sol}\,[HA(sol)]/[A^-(sol)]. \qquad (123)$$

Substitution of Eq. (122) into Eq. (116) gives

$$E_{Pt} = 0.059 \log (K_a)_{sol} + 0.059 \log [HA(sol)]/[A^-(sol)]$$
$$-(0.059/2) \log P_{H_2}. \qquad (124)$$

This equation in conjunction with voltammetric measurements of half-wave potentials ($E_{1/2}$) for the reduction of Bronsted acids at a platinum electrode in any solvent permits the evaluation of (K_a) sol [pK_a(sol)],

$$E_{1/2} = 0.059 \log (K_a)_{sol} + 0.059 \log (\kappa_{A^-}\kappa_{H_2}/\kappa_{HA})$$
$$= -0.059\, pK_a + \varepsilon, \qquad (125)$$

where κ_{A^-}, κ_{H_2}, and κ_{HA} are parameters that relate to diffusion coefficients, activity coefficients, and P_{H_2} for a given experimental system. Because solvent has such a major and selective effect on the activity of H_3^+O(aq), the differences between concentration-based dissociation constants (pK) and activity-based constants (pK_a) are dramatic. For example, phenol in MeCN (pK, 26.6; pK_a, 16.0) and in Me_2SO (pK, 16.4; pK_a, 20.8) exhibits a reversal; it dissociates more in Me_2SO than in MeCN, but is more acidic [greater H_3^+O(aq) activity] in MeCN. Other examples include (Et_3NH)Cl in MeCN (pK, 18.5; pK_a, 10.0) and in Me_2SO (pK, 10.5; pK_a, 12.7) (again, greater dissociation in Me_2SO and greater acidity in MeCN); PhC(O)OH in MeCN (pK, 20.7; pK_a, 7.9), in Me_2SO (pK, 11.1; pK_a 13.6), in DMF (pK, 11.6; pK_a, 11.5), and in H_2O (pK, 4.2; pK_a, 3.2); and H_2O in MeCN (pK_a, 30.4) and in Me_2SO (pK, 31.4; pK_a, 36.7).

3. Oxidation of Dissolved Dihydrogen (H_2)

Molecular hydrogen (H_2; ΔH_{DBE}, 104 kcal mol^{-1}) is resistant to electrochemical oxidation at inert electrodes (glassy carbon or passivated metals; Ni, Au, Hg, Cu). At passivated Pt and Pd, dissolved H_2 only exhibits broad, diffuse, anodic voltammetric peaks with irreproducible peak currents that are not proportional to the partial pressure of dissolved H_2 (P_{H_2}). However, with freshly preanodized Pt and Pd electrodes, well-defined oxidation peaks for H_2 are obtained, which have peak currents that are proportional to P_{H_2}. The surface conditioning produces a fresh reactive metal-oxide surface [$Pt^{II}(OH)_2(s)$], which upon exposure to H_2 becomes an oxide-free metal surface (Pt*). In turn, the clean surface reacts with a second H_2 to form two Pt–H bonds,

$$2\,Pt^*(s) + H_2(P_{H_2}) \xrightleftharpoons{fast} 2\,Pt\text{–}H(s) \qquad K_{eq}, \sim 1 \text{ atm}^{-1}. \qquad (126)$$

The value of K_{eq} is estimated on the basis of the Pt–H bond-formation energy from metallic platinum [$\Delta(-\Delta G_{BF})$, 48 kcal mol^{-1}] and the dissociative bond energy for H_2 (ΔG_{DBE}, ~96 kcal mol^{-1}).

This activated platinum electrode system is the equivalent of an NHE and, therefore, conforms to the Nernst expression of Eq. (116). Rearrangement gives

$$\log P_{H_2} = -E_{Pt}\,(2/0.059) + \log\left[H_3^+O(aq)\right]^2, \qquad (127)$$

which responds to the overall reaction

$$H_2\left(P_{H_2}\right) + 2\,Pt^*\,(s) \underset{}{\overset{K_{eq}}{\rightleftharpoons}} 2\,Pt\text{–}H\,(s) \underset{H_2O}{\overset{-2e^-}{\rightleftharpoons}} 2\,H_3^+O(aq) + 2\,Pt\,(s). \qquad (128)$$

Although there is a long-standing tradition to write an electron-transfer sequence with electron removal (ionization) from the hydrogen atom to produce a proton [Pt–H (s) → Pt + $H^+ + e^-$] followed by hydration to give the observed hydronium ion [$H^+ + H_2O \rightarrow H_3^+O(aq)$], the ionization potential for a free hydrogen atom (H·) is 13.6 eV and even greater for bound hydrogen [Pt–H (s)]. Electron removal from an aqueous solution at pH 0 with a GCE (free of H_2 and Pt–H) occurs via the lowest energy path, which is oxidation of the solvent

$$H_2O \xrightarrow{-e^-} \left[H_2^+O\cdot\right] \xrightarrow{H_2O} H_3^+O + HO\cdot \qquad E^\circ,\ +2.72\text{ V vs NHE}. \qquad (129)$$

In the presence of Pt–H (s) (and H_2), this process is facilitated via direct formation of an H–OH bond,

$$H_2O + Pt\text{–}H\,(s) \underset{H_2O}{\overset{-e^-}{\rightleftharpoons}} Pt\,(s) + H_3^+O + HO\text{–}H \qquad E^\circ,\ 0.00\text{ V}$$

$$\Delta(-\Delta G_{BF}) = (2.72 - 0.00)\,23.1\text{ kcal (eV)}^{-1} = 63\text{ kcal mol}^{-1} \qquad (130)$$

Addition of the Pt–H (s) bond energy [$\Delta(-\Delta G_{BF}) = 48$ kcal mol^{-1}] to the differential bond-formation energy for H–OH gives an electrochemical measure of the HO–H bond energy [$-\Delta G_{BF} = 111$ kcal mol^{-1} (literature value, 111 kcal mol^{-1})].

B. Dioxygen Species (·O_2·, HOO·/O_2^-·, HOOH)

Although the electrochemistry of hydronium ion and molecular hydrogen is fundamental, the electrochemical characterization of oxygen species (·O_2·, HOO·, O_2^-·, HOOH, HOO^-, ·O·, HO·, O^-·, H_2O, HO^-, O_3, O_3^-·) and of the oxygen component of molecules (e.g., M_xO_y, oxy anions and radicals, quinones) is its most important and unique application. In general, electrochemical measurements provide the only direct means for the evaluation of the electron-transfer thermodynamics of oxygen species and oxygen-containing molecules. Also, amperometric sensors for ·O_2· are the most common analytical methodology for its assay in blood, gas streams, biotreaters, and process streams.

1. Molecular Oxygen

The ground state of molecular oxygen (·O_2·, $^3\sum_g^-$) has two unpaired electrons in degenerate $2\pi_g$ orbitals and is referred to as *dioxygen* by most contemporary biologists and biochemists. When dioxygen is reduced by electron transfer, a series of intermediate basic dioxygen and monooxygen species are produced that may take up one or two hydronium ions (H_3^+O) from the media (O_2^-·, HOO·, HOO^-, HOOH, ·O^-, HO·, HO^-, and H_2O). The thermodynamics for the various reduction steps in aqueous solutions have been evaluated by numerous techniques, but all are fundamentally based on the calorimetry associated with the reaction

$$\cdot O_2\cdot\,(g) + 2\,H_2\,(g) \longrightarrow 2\,H_2O\,(l) \qquad -\Delta G^\circ = nE^\circ_{cell}F. \qquad (131)$$

If redox potentials are relative to the normal hydrogen electrode,

$$2\,H_3^+O + 2e^- \longrightarrow H_2\,(g) \qquad E^\circ_{H_3^+O/H_2} \equiv 0.000, \qquad (132)$$

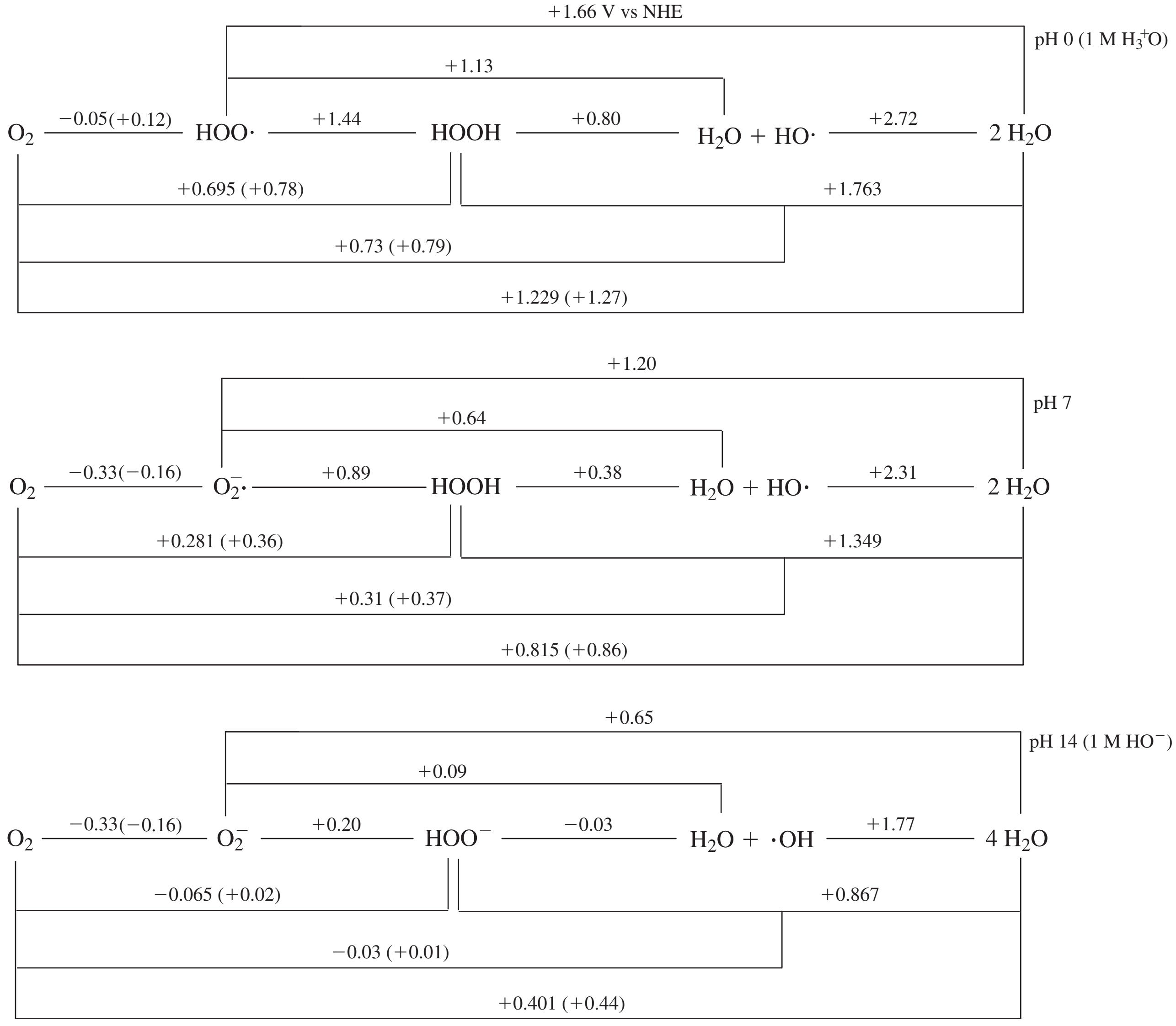
then the standard redox potential for the four-electron reduction of dioxygen ($E^\circ_{O_2/H_2O}$) can be calculated from the calorimetric data for the reaction of Eq. (131) under standard-state conditions,

$$\cdot O_2\cdot\,(g) + 4\,H_3^+O(aq) + 4e^- \longrightarrow 2\,H_2O\,(l)$$

$$E^\circ_{O_2/H_2O} = E^\circ_{cell} - E^\circ_{H_3^+O/H_2} = +1.229\text{ V vs NHE}. \qquad (133)$$

The reduction potentials for ·O_2· and various intermediate species in H_2O at pH 0, 7, and 14 are summarized in Table IV; similar data for ·O_2· in MeCN at pH −8.8, 10.0, and 30.4 are presented in Table V. For those couples that involve dioxygen itself, formal potentials are given in parentheses for ·O_2· at unit molarity (~10^3 atm; [·O_2·] ≈1 mM at 1 atm partial pressure). The reduction manifolds for ·O_2· (Tables IV and V) indicate that the limiting step (in terms of reduction potential) is the first electron transfer to ·O_2· and that an electron source adequate for the reduction of ·O_2· will produce all of the other reduced forms of dioxygen (O_2^-·, HOO·, HOOH, HOO^-, HO·, H_2O, HO^-) via reduction, hydrolysis, and disproportionation steps. Thus, the most effective means to activate ·O_2· is the addition of an electron (or hydrogen atom; $H_3^+O + e^- \rightarrow$ H·), which results in significant fluxes of several reactive oxygen species.

TABLE IV Standard Reduction Potentials for Dioxygen Species in Water [O_2, 1 atm (1 mM)] (Formal Potentials for 1 M O_2)



2. Aprotic Media

In the absence of proton sources dioxygen is reversibly reduced to superoxide ion,

$$\cdot O_2\cdot + e^- \rightleftharpoons O_2^-\cdot \qquad E^{\circ\prime}, -0.64 \text{ V vs NHE.} \quad (134)$$

The second reduction is an irreversible one-electron process,

$$O_2^-\cdot + e^- \xrightarrow{H_2O} HOO^- + HO^- \qquad E_{p,c}, -1.8\text{ V} \quad (135)$$

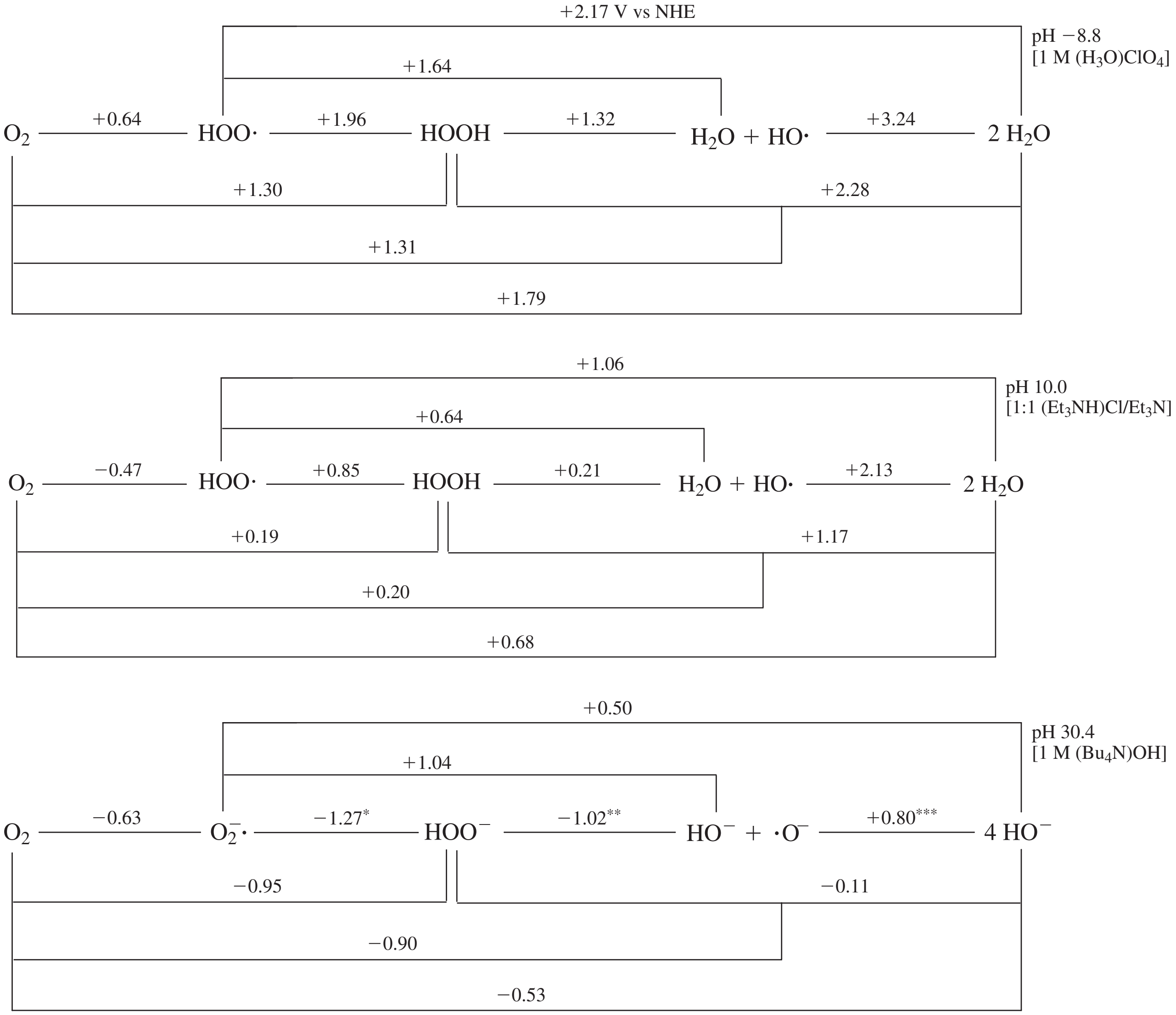
$$HO^- + HOO^- + \cdot O_2\cdot \rightleftharpoons 2\,O_2^-\cdot + H_2O$$
$$K, 10^{-9}\text{ M}^{-1}\text{ atm}^{-1}. \quad (136)$$

The reduction potential for the $\cdot O_2\cdot/O_2^-\cdot$ couple shifts to more negative values as the solvating properties of the media decrease. The heat of hydration ($-\Delta H_{aq}$) for gaseous $O_2^-\cdot$ is 100 kcal mol^{-1}, which is consistent with the unique strong solvation of anions by water. Hence, if the $E^{\circ\prime}_{O_2/O_2^-\cdot}$ values for the $\cdot O_2\cdot/O_2^-\cdot$ couple are affected primarily by the degree of solvation of $O_2^-\cdot$ (that is, the solvation energy for $\cdot O_2\cdot$ is assumed to be small and about the same for different solvents), then the relative solvation energies for $O_2^-\cdot$ are $H_2O \gg Me_2SO > DMF > py \sim$ MeCN.

3. Hydrogen Peroxide (HOOH, H_2O_2)

The electrochemical reduction of HOOH in pyridine yields a superoxide ion. That a reduction process generates a species with a higher oxidation state than HOOH

TABLE V Formal Reduction Potentials for 1 M Dioxygen Species in Acetonitrile [O_2 at 1 atm (8.1 mM)]



*($O_2^-\cdot \xrightarrow{-1.51\ V}$ HOOH).

**(HOOH $\xrightarrow{-0.90\ V}$ HO^- + HO·).

***(HO· $\xrightarrow{+0.92\ V}$ HO^-).

is surprising. The cyclic voltammogram for the reduction of HOOH in pyridine at a platinum working electrode exhibits a broad cathodic peak at −0.95 V vs NHE and anodic peaks for the reverse scan at −0.50 and −0.15 V vs NHE. The latter is characteristic of electrolytically generated H_2. Controlled-potential reduction of HOOH in pyridine at −1.0 V vs NHE (with argon degassing) results in a solution that exhibits an anodic cyclic voltammogram that is characteristic of $O_2^-\cdot$. ESR studies of the reduced solution confirm the presence of a superoxide ion. The products and the observed electron stoichiometries for the electrochemical reduction of HOOH are consistent with a mechanism in which the primary step is a one-electron transfer,

$$\text{HOOH} + e^- \longrightarrow \text{HOO}^- + 1/2\,\text{H}_2 \qquad E^{\circ\prime}, -1.0\ \text{V vs NHE} \qquad (137)$$

followed by a chemical reaction with another hydrogen peroxide molecule.

$$HOOH + HOO^- \rightleftharpoons [HOOOH] + HO^-$$

$$\xrightarrow[k, 3\times10^9\ M^{-1}\ s^{-1}]{py} 1/n\,[\cdot pyOH]_n + H_2O + O_2^-\cdot \quad (138)$$

The resultant HO· is trapped by the pyridine solvent to yield a stable solution of $O_2^-\cdot$.

In acetonitrile there is no evidence of a superoxide ion, either by ESR or by cyclic voltammetry, from the electrochemical reduction of HOOH. This can be explained by the slow rate of reaction of HO· with CH_3CN (k, $4 \times 10^6\ M^{-1}\ s^{-1}$), which favors direct disproportionation.

$$[HOOOH] + HO^- \rightarrow H_2O + [HOOO^-] \rightarrow HO^- + \cdot O_2\cdot \quad (139)$$

The reduction of hydrogen peroxide in aqueous solution appears to be analogous to that in CH_3CN, with the mechanism represented by the reaction of Eq. (137) followed by the reactions of Eqs. (138) and (139). Thus, the reduction of HOOH yields H_2 and HOO^-, initially, in a one-electron step. The final products are the result of the reaction of HOO^- and HOOH and are analogous to those for the base-induced decomposition of HOOH.

In acetonitrile HOOH is oxidized to $\cdot O_2\cdot$ via an electron-transfer/chemical/electron-transfer (ECE) mechanism,

$$HOOH \xrightarrow{-e^-} [HOOH]^{+}\cdot \xrightarrow{H_2O} HOO\cdot + H_3^+O \qquad E^{\circ\prime},\ +2.0\ V\ vs\ NHE \quad (140)$$

$$HOO\cdot \xrightarrow[H_2O]{-e^-} \cdot O_2\cdot + H_3^+O \quad E^{\circ\prime},\ +0.6\ V\ vs\ NHE. \quad (141)$$

Although HOO^- reacts rapidly with most organic solvents, it persists long enough in pyridine to permit its electrochemical oxidation via a similar ECE mechanism.

$$HOO^- \xrightarrow{-e^-} [HOO\cdot] \xrightarrow{HO^-} H_2O + O_2^-\cdot \qquad E^{\circ},\ -0.37\ V\ vs\ NHE \quad (142)$$

$$O_2^-\cdot \xrightarrow{-e^-} \cdot O_2\cdot \qquad E^{\circ},\ -0.66\ V\ vs\ NHE \quad (143)$$

C. Nonmetals

Although hydronium ion (H_3^+O) and dioxygen ($\cdot O_2\cdot$) are the most studied of the molecules and ions without metal atoms, several of the molecules that contain sulfur, nitrogen, or carbon also are electroactive. The results for representative examples illustrate the utility of electrochemical measurements for the evaluation of the redox thermodynamics and bond energies for nonmetals. In particular, the electrochemistry for several sulfur compounds [S_8, SO_2], nitrogen compounds [·NO, hydrazines (RNHNHR')], and carbon compounds (CO_2, CO) is summarized and interpreted.

Some of the considerations for electron-transfer processes that have been discussed are fundamental to the electrochemistry of these examples. Thus, reductive processes always involve the most electrophilic (acidic, positive-charge density) center (substrate or substrate/matrix combination) that produces the least basic (nucleophilic) product. Under acidic conditions the primary reactant often is the hydronium ion (H_3^+O) to give a hydrogen atom that couples with the substrate via covalent bond formation, e.g.,

$$PhN{=}NPh + 2\,H_3^+O + 2e^- \xrightarrow[Me_2SO]{} PhNHNHPh + 2\,H_2O \quad E_{p,c},\ -0.5\ V\ vs\ SCE. \quad (144)$$

In contrast, oxidations always involve the most nucleophilic (basic, negative-charge density) center (substrate or substrate/base combination) that produces the least acidic (electrophilic) product. Under basic conditions (or neutral aqueous solutions) the primary reactant often is the base [B^-; HO^-, $HOC(O)O^-$, AcO^-, PhO^-] to give an oxyl radical (e.g., HO·) that adds to the substrate or, more often, abstracts a hydrogen atom, e.g.,

$$PhNHNHPh + 2HO^- \xrightarrow[MeCN]{-2e^-} PhN{=}NPh + 2\,H_2O \qquad E_{p,a},\ -1.1\ V\ vs\ SCE. \quad (145)$$

1. Elemental Sulfur (S_8)

Although the cyclic voltammogram of S_8 (in dimethyl sulfoxide at a gold electrode) appears to be a simple sequence of two two-electron-per-S_8, pseudo-reversible reductions, controlled-potential coulometry at −0.7 V vs SCE indicates that 2.7 electrons per S_8 are consumed to give S_6^{2-}. Controlled-potential coulometry at −1.5 V vs SCE consumes 4.0 electrons per S_8 to produce two S_4^{2-} ions. On the basis of these results and associated spectroscopic studies, the reduction sequence in Me_2SO involves two ECE steps with ring opening and chain breaking the two chemical steps,

$$c\text{-}S_8 + e^- \xrightarrow[E_{p,c}, -0.60\ V\ vs\ SCE]{} \cdot S_8^- \xrightarrow{e^-} {}^-S_8^- \xrightarrow[E_{p,c},\ -1.29\ V]{e^-} S_4^{2-} + S_4^-\cdot \xrightarrow{e^-} S_4^{2-} \quad (146)$$

The appearance of the bright blue solution ($S_3^-\cdot$) during the course of electrolysis is dramatic;

$$3S_8 + 8e^- \xrightarrow{-0.6\ \text{V vs SCE}} 8S_3^{-}\cdot \quad \rightleftarrows \quad 4S_6^{2-} \quad (K_D,\ 1.1 \times 10^2\ \text{M}^{-1}) \qquad (147)$$

2. Sulfur Dioxide (SO_2)

In aprotic solvents SO_2 undergoes a reversible one-electron reduction,

$$SO_2 + e^- \underset{\text{DMF}}{\rightleftarrows} SO_2^{-}\cdot \quad E_{1/2},\ -0.75\ \text{V vs SCE}. \qquad (148)$$

However, the product species interacts with excess SO_2 to give a blue complex that is significantly more difficult to oxidize.

$$SO_2^{-}\cdot + 2\,SO_2 \longrightarrow (SO_2)_2\,\dot{S}(O)O^- \xrightarrow{-e^-,\ E_{p,a},\ -0.25\ \text{V vs SCE}} 3SO_2 \qquad (149)$$

Complete electrolysis yields $SO_2^{-}\cdot$, which dimerizes to dithionite ion (colorless),

$$2SO_2^{-}\cdot \underset{\text{DMF}}{\rightleftarrows} {}^{-}O_2SSO_2^{-} \quad K_D,\ 2.4 \times 10^1\ \text{M}^{-1}. \qquad (150)$$

Because SO_2 is electrophilic (acidic), it is extremely resistant to direct electron-transfer oxidation. However, in aqueous solutions at pH 1, SO_2 facilitates the oxidation of H_2O [$2\,H_2O \rightarrow HO\cdot + H_3^+O + e^-$; $E^{\circ\prime}$, +2.42 V vs SCE (pH 1)] at gold electrodes via covalent-bond formation to give sulfuric acid [$(HO)_2S(O)_2$] in an ECE process. The peak current is proportional to the SO_2 concentration and its diffusion coefficient, which makes this anodic process suitable as a voltammetric monitor for dissolved SO_2 or gas-phase SO_2 via a gas-permeable membrane.

3. Nitric Oxide (·NO)

Reduction of the oxides of nitrogen (·NO, $\cdot NO_2$, and N_2O) usually involves the addition of hydrogen atoms that are electrogenerated. The use of a mercury electrode inhibits the reduction of H_3^+O to H_2 ($E^{\circ\prime}$, −2.2 V vs SCE at pH 5), but allows formation of H· when it couples with a substrate via covalent-bond formation,

$$\cdot N{=}O + H_3^+O\ (\text{pH } 5.0) + e^- \longrightarrow [NO{-}H] \quad E_{1/2},\ -0.9\ \text{V vs SCE}$$
$$[NO{-}H] \xrightarrow{2X} [HON{=}NOH] \longrightarrow N{\equiv}N{=}O + H_2O\ \ (N_2O) \qquad (151)$$

4. Hydrazines and Amines

These substrates are directly oxidized in a base-free matrix (Me_2SO or MeCN) at platinum or glassy-carbon electrodes with the potential primarily determined by the *R*N-H bond energy and secondarily by the basicity of the substrate,

$$2\ \text{PhNHNHPh}\ (R\text{NH}) \xrightarrow[E_{p,a},\ +0.6\ \text{V vs SCE}]{-e^-} \text{Ph}\dot{\text{N}}\text{NHPh} + R\text{NH}_2^+$$
$$\text{Ph}\dot{\text{N}}\text{NHPh} \xrightarrow[R\text{NH}]{-e^-} \text{PhN}{=}\text{NPh} + R\text{NH}_2^+ \qquad (152)$$

$$2\ \text{PhNH}_2 \xrightarrow[E_{p,a},\ +0.9\ \text{V vs SCE}]{-e^-} \text{Ph}\dot{\text{N}}\text{H} + \text{PhNH}_3^+$$
$$\text{Ph}\dot{\text{N}}\text{H} \xrightarrow[-e^-]{2\ \text{PhNH}_2} \text{PhNHNHPh} + \text{PhNH}_3^+$$
$$\text{PhNHNHPh} \xrightarrow[-2e^-]{2\ \text{PhNH}_2} \text{PhN}{=}\text{NPh} + 2\ \text{PhNH}_3^+. \qquad (153)$$

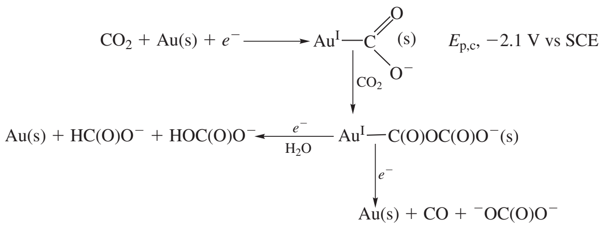
In contrast, when a hydroxide ion HO^- is present it is more easily oxidized than the amine substrates. In MeCN, in the absence of substrate, HO^- is oxidized at +0.7–0.9 V vs SCE. However, with hydrazines and amines present, the N—H bonds are homolytically cleaved by the HO· product of HO^- oxidation. The latter's oxidation potential is shifted by the difference in the HO—H and *R*N—*H* bond energies ($-\Delta G_{BF}$). Thus, the oxidation of PhNHNHPh is shifted by −1.7 V when HO^- becomes the electron-transfer mediator; with $PhNH_2$ the shift is by −1.1 V.

$$\text{PhNHNHPh} + HO^- \xrightarrow[E_{p,a},\ -1.1\ \text{V vs SCE}]{-e^-} \text{Ph}\dot{\text{N}}\text{NHPh} + H_2O$$
$$\text{Ph}\dot{\text{N}}\text{NHPh} \xrightarrow[HO^-]{-e^-} \text{PhN}{=}\text{NPh} + H_2O \qquad (154)$$

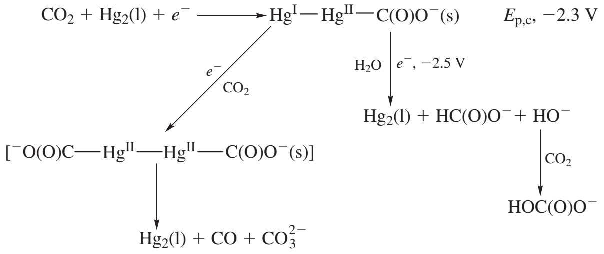
$$\text{PhNH}_2 + HO^- \xrightarrow[-0.2\ \text{V vs SCE}]{-e^-} \text{Ph}\dot{\text{N}}\text{H} + H_2O$$
$$\text{Ph}\dot{\text{N}}\text{H} \xrightarrow[-e^-]{HO^-,\ \text{PhNH}_2} \text{PhNHNHPh} + H_2O \qquad (155)$$

5. Carbon Dioxide (CO_2)

Whereas the reduction of CO_2 at a gold electrode is a one-electron per CO_2 process on a voltammetric time scale, at mercury it is a sequential two-electron process. In both cases the overall reduction is two electrons per CO_2. The products for anhydrous conditions are CO_3^{2-} and CO and, with H_2O present, $HOC(O)O^-$ and $HC(O)O^-$,



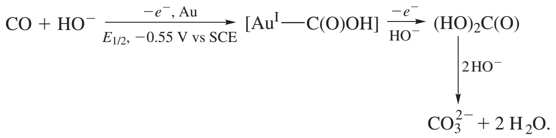
(156)



(157)

6. Carbon Monoxide (CO)

At an activated gold electrode in alkaline (0.01 M NaOH) aqueous solution CO is oxidized to CO_3^{2-} via a HO^--centered ECE process,



(158)

Again, the activated gold surface stabilizes the carbon radical intermediate [·C(O)OH] and facilitates the second electron-transfer oxidation of HO^- to HO· via coupling to the carbon radical.

D. Organic Molecules

Most organic reactions are Lewis acid/base processes that involve the interaction of a nucleophilic center with an electrophilic center. Because electrochemistry provides the ultimate nucleophile via the electrons at the cathode surface and the ultimate electrophile via the electron holes at the anode surface, it is the ideal methodology for the characterization of the electrophilicity and nucleophilicity of molecules. Thus, the carbon centers of saturated hydrocarbons (e.g., CH_4) are resistant to electrochemical reduction and oxidation because of their inert nature (all valence electrons are stabilized in sigma bonds; an absence of any Lewis acid/base character). However, organic molecules with electrophilic components [e.g., alkyl-, aryl-, and acyl- halides; carbonyl groups; unsaturated and aromatic hydrocarbons; nitro groups; Brønsted acids H*A*] react with the ultimate nucleophile (the electron) directly or via its reaction product with hydronium ion (H_3^+O) or Brønsted acids (H*A*).

For those electrophiles (*El*) that undergo direct electron-transfer reduction at an inert electrode (glassy carbon) [with the electron occupying the Lowest-Unoccupied-Molecular-Orbital (LUMO)], the reduction potential (E_{red}) is a measure of their electron affinity and electrophilicity [relative to that for H_3^+O (−2.10 V vs NHE in aqueous media)] (the more positive the potential, the more electrophilic the molecule).

$$El + e^- \rightleftarrows El^-\cdot \quad E_{red} \qquad (159)$$

Often, the solution matrix contains Lewis acids (e.g., H_3^+O) that are more electrophilic than the substrate molecule and, in combination, are even more electrophilic,

$$El + H_3O^+ + e^- \xrightarrow{\;-H_2O\;} [HEl] \xrightarrow[-H_2O]{e^-,\,H_3O^+} HElH \quad (E_{red})_{EC.} \qquad (160)$$

The first-formed intermediate (with an unpaired electron) in combination with a second Lewis acid molecule has even greater electron affinity and is reduced at a more positive potential to give a voltammogram that appears to be the result of an irreversible two-electron reduction process. In most cases it is an ECEC process in which each electron transfer (E part of the ECEC mechanism) to the Lewis acid (H_3^+O) is reversible to give a product (H·) that forms a covalent bond with the substrate (H–*El*) (the C part of the mechanism).

Conversely, nucleophilic molecules (*Nu*) [Lewis bases; e.g., catechols, hydroquinones, phenols, alcohols, and thiols (and their anions); aromatic hydrocarbons and amines (benzene, toluene, pyridine, bipyridine, etc.)] can be oxidized (1) by direct electron-transfer oxidation [Eq. (161)] [with the electron coming from the Highest-Occupied-Molecular-Orbital (HOMO)] or (2) by coupling with the oxidation product of H_2O (or HO^-), hydroxyl radical (HO·) [Eq. (162)].

$$Nu\!: \overset{-e^-}{\rightleftarrows} Nu^+\cdot \quad E_{ox} \qquad (161)$$

The potential (E_{ox}) for those nucleophiles that undergo direct electron-transfer oxidation at an inert electrode is a quantitative measure of their nucleophilicity (the more negative the potential the more nucleophilic the molecule; see Table VI for representative values). In many cases water in the solvent matrix (or as the solvent) is more nucleophilic than the substrate molecule and, in combination, is even easier to oxidize, which often results in an ECEC oxidation process.

TABLE VI Nucleophilicity and Electrophilicity of Molecules and Ions (Strongest or Most Reactive at Top of Listing)

	$(E_{1/2})_{ox}$, V vs NHE			$(E_{1/2})_{red}$, V vs NHE	
Nucleophile	**(MeCN)**	**(H_2O)**	**Electrophile**	**(MeCN)**	**(H_2O)**
e^-_{aq}	−3.9	−2.9	$H_2O^{+}\cdot$	+3.2	+2.7
K·	−2.9	−2.9	$Ph^{+}\cdot CH_2OH$	+2.2	
Na·	−2.1	−2.7	$Au^{I}(OH_2)_6^{+}$	+1.6	+1.8
			$(Cl_8TPP^{+}\cdot)Fe^{IV}{=}O$ (Compound I)	+1.5	
Li·	−2.0		$Fe^{III}(bpy)_3^{3+}$	+1.3	+1.1
$(TPP^{-}\cdot)Co^{-}$	−1.7		HO·	+0.9	+1.9
H·	−1.6	−2.1	$Fe^{III}(PA)_3$	+0.4	
$(TPP^{-}\cdot)Fe^{-}$	−1.4		$(TPP)Fe^{III}(py)_2^{+}$	+0.4	
$(TPP)Fe^{-}$	−0.8		$(TPP)Fe^{III}Cl$	+0.2	
$(TPP)Co^{-}$	−0.6		MV^{2+}	−0.2	
$O_2^{-}\cdot$	−0.7	−0.2	$(Cl_8TPP)Fe^{IV}{=}O$ (Compound II)	−0.3	
$PhCH_2S^{-}$	0.0		AQ (Anthraquinone)	−0.6	
HOO^{-}	0.0	+0.7	O_2	−0.7	−0.2
PhO^{-}	+0.3		CCl_4	−0.9	
Me_3N	+0.7		$PhCH_2Br$	−1.4	
HO^{-}	+0.9	+1.9	$PhCl_6$	−1.4	
$MeC(O)O^{-}$	+1.5		*t*-BuI	−1.5	
PhOH	+1.7		H_3O^{+}	−1.6	−2.1
pyridine	+2.0		$PhCH_2Cl$	−1.7	
Cl^{-}	+2.2	+2.4	*n*-BuI	−1.9	
HOOH	+2.3	+1.0	*c*-$C_6H_{11}Br$	−1.9	
H_2O	+3.0	+2.3	*t*-BuBr	−2.0	
			n-BuBr	−2.2	
			n-BuCl	−2.5	
			PhCl	−2.6	
			H_2O	−3.9	−2.9

$$Nu{:} + 2\,H_2O \xrightarrow[\;\;H_3O^{+}]{-e^{-}} [HO{-}Nu\cdot] \xrightarrow[\;\;H_3O^{+}]{-e^{-},\,2\,H_2O} Nu(OH)_2 \qquad (E_{ox})_{EC} \tag{162}$$

1. Alkyl- and Aryl-Halides

Because the halo-groups of organic molecules have large electronegativities and electron affinities, all halo-carbon molecules are electrophilic. Their electrochemical reduction potential is a measure of their electrophilicity (and electron affinity), which is illustrated in Fig. 6 for hexachlorobenzene (C_6Cl_6), 1,2,3,4-tetrachlorobenzene (1,2,3,4-$C_6H_2Cl_4$), and *n*-butyl iodide (*n*-BuI). Table VII summarizes the reduction potentials for several alkyl-halides and aryl-chlorides.

In the absence of other Lewis acids, alkyl-halides undergo direct electron addition at the electrode surface with subsequent stimulated electron transfer (a) to a second substrate and coupling (*R*–*R*) or (b) to a Brønsted acid (H*A*; H_2O) to replace the C–X bond with a C–H bond. For example,

$$n\text{-BuI} + e^{-} \longrightarrow [n\text{-Bu}\dot{\text{I}}^{-}] \qquad E_{p,c},\ -2.05\text{ V vs SCE}$$
$$\text{(a)}\ \xrightarrow{n\text{-BuI},\,e^{-}} n\text{-Bu–Bu-}n + 2\,I^{-}$$
$$\text{(b)}\ \xrightarrow{H_2O,\,e^{-}} n\text{-BuH} + I^{-} + HO^{-} \tag{163}$$

In both cases the overall process is an irreversible two-electron reduction via either (a) an EE path or (b) an ECEC path; the first electron transfer is the most difficult and depends on the substrates electrophilicity. In the presence of an hydronium ion, the primary electron transfer will be to the most electrophilic center, e.g.,

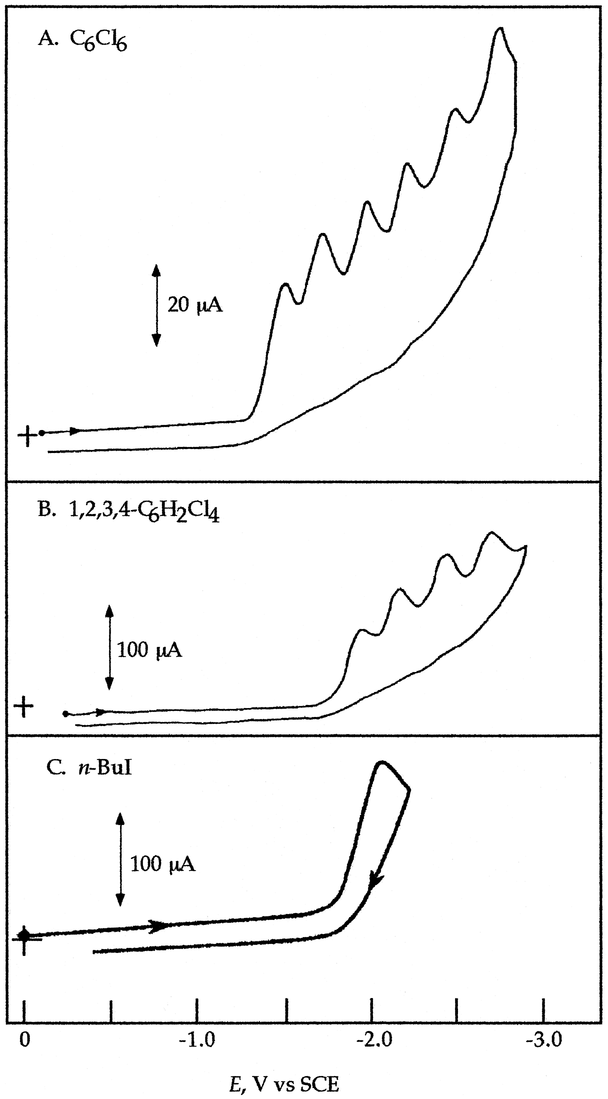


FIGURE 6 Cyclic voltammograms for chlorinated aromatic molecules and *n*-butyl iodide in dimethylformamide (0.1 M TEAP) at a glassy-carbon electrode (area, 0.062 cm^2): (a) 1.1 mM hexachlorobenzene (C_6Cl_6); (b) 2.3 mM 1,2,3,4-$C_6H_2Cl_4$; (c) 20 mM *n*-BuI.

$$H_3O^+ + n\text{-BuI} + e^- \xrightarrow{\;-H_2O\;} [n\text{-Bu}\dot{I}H] \xrightarrow[-H_2O]{H_3O^+,\, e^-} n\text{-BuH} + HI. \qquad (164)$$

Thus, the reduction of n-BuI is the equivalent of the addition of two hydrogen atoms [H·] (generated via the electrochemical reduction of the two hydronium ions).

With aryl-chlorides (e.g., $PhCl_6$) a similar EEC process occurs as each chlorine atom is replaced with a hydrogen atom (Fig. 6 and Table VII).

$$PhCl_6 + e^- \longrightarrow [PhCl_6^{-}\cdot] \xrightarrow{e^-,\, H_2O} HPhCl_5 + Cl^- + HO^- \qquad (165)$$

Hence, the $PhCl_6$ exhibits six irreversible two-electron reductions (each product species less electrophilic than its precursor) to yield at −2.8 V vs SCE benzene (PhH); an overall 12-electron process.

$$PhCl_6 + 6\,H_2O + 12e^- \longrightarrow PhH + 6\,Cl^- + 6\,HO^- \qquad (166)$$

Although such electrolyses are done in aprotic solvents (e.g., DMF, DMSO, MeCN), even the most rigorously dried solvent contains 3–20 mM H_2O (50–350 ppm). If the solvent has a degree of Brønsted acidity (e.g., alcohols and ketones), then it can serve as a source of hydrogen atoms.

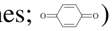
TABLE VII Redox Potential for Alkyl Halides (*RX*) and Aryl Chlorides (*Ar*Cl_x in Dimethyl Formamide at a Glassy Carbon Electrode[a]

Alkyl halides			
	$E_{p,c}$,[b] V vs SCE[c]		
RX	**I**	**Br**	**Cl**
CH_3–	−2.10		
n-C_4H_9–	−2.05	−2.41	
sec-C_4H_9–	−1.92	−2.35	
t-C_4H_9–	−1.78	−2.25	
c-C_6H_{11}–	−2.03	−2.48	
$PhCH_2$–		−1.68	−1.90
ClH_2C–			−2.05
Cl_2HC–			−1.99
Cl_3C–		−1.13	
FCl_2C–			−1.71
$(F_3C)Cl_2C$–			−1.31
$PhCl_2C$–			−1.47
[(p-ClPh)$_2$HC]Cl_2C– (DDT)			−1.50
[(p-ClPh)$_2$FC]Cl_2C– (F-DDT)			−1.44

Alkyl halides			
***Ar*Cl$_x$**	**$E_{p,c}$, V vs SCE**	***Ar*Cl$_x$**	**$E_{p,c}$, V vs SCE**
PhCl(C_6H_5Cl)	−2.7	$PhCl_5$	−1.6
1,2-$PhCl_2$	−2.5	$PhCl_6$	−1.4
1,2,3-$PhCl_3$	−2.2	Cl_5Ph-$PhCl_5$	−1.5
1,2,3,4-$PhCl_4$	−1.9		

[a] mM solutions in DMF [0.1 M (Et_4N)ClO_4].
[b] $E_{p,c}$, the reduction-peak potential.
[c] Saturated calomel electrode (SCE) vs NHE, +2.44 V.

2. Quinones, Semiquinones, and Catechols

All molecules with unsaturated bonds (olefins, acetylenes, aromatics, carbonyls, quinones, etc.) have a degree of electrophilicity and electron affinity. Within a class, the extent of conjugation increases the electron affinity (reduction of benzene occurs at a less negative potential than 1-butene), and the presence of unsaturated carbon-oxygen (carbonyl) functions within a conjugated system (e.g., quinones; ) also enhances the electron affinity of the molecule.

The cyclic voltammogram for 3,5-di-*tert*-butyl-*o*-quinone (3,5-DBTQ) has a first reduction that is a reversible one-electron process, followed by a second one-electron reduction, which can be reversible in rigorously anhydrous media to give catechol dianion (3,5-DTBC^{2-}).

$$\text{3,5-DTBQ} + e^- \rightleftharpoons \text{3,5-DTBSQ}^-\cdot \overset{e^-}{\rightleftharpoons} \text{3,5-DTBC}^{2-} \quad (167)$$

Although DTBC^{2-} is a strong base that is hydrolyzed by residual H_2O,

$$\text{DTBC}^{2-} + H_2O \rightleftharpoons \text{DTBCH}^- + HO^-, \quad (168)$$

it is also a strong reductant that can reduce H_2O in DMF.

$$\text{DTBC}^{2-} + H_2O \longrightarrow \text{DTBSQ}^-\cdot + 1/2\,H_2 + HO^- \quad (169)$$

In many cases, the second reduction step of quinones [Eq. (167)] is irreversible and due to the facilitated reduction of residual H_2O.

$$\text{3,5-DTBSQ}^-\cdot + H_2O + e^- \longrightarrow \text{3,5-DTBCH}^- + HO^- \quad (170)$$

In the presence of hydronium ions (H_3^+O), the reduction of the quinones is an irreversible two-electron process (ECEC), with the first step being the more difficult (requiring the more negative potential).

$$\text{3,5-DTBQ} + H_3O^+ + e^- \xrightarrow{-H_2O} [\text{3,5-DTBS}\dot{\text{Q}}\text{H}] \xrightarrow[-H_2O]{e^-,\ H_3O^+} \text{3,5-DTBCH}_2 \quad (171)$$

The oxidation potentials for the fully reduced forms of quinones (catechols; H_2Cat, $HCat^-$, and hydroquinones; H_2Q, HQ^-) indicate that the first electron removal is the most difficult [e.g., 3,5-DTBCH$^-$ in DMF] in an irreversible two-electron oxidation via an ECE mechanism.

$$HCat^- \xrightarrow{-e^-} HSQ\cdot \xrightarrow[HCat^-]{-e^-} Q + H_2Cat \qquad E_{p,a},\ +0.14\text{ V vs SCE} \quad (172)$$

Neutral catechols (H_2Cat) and hydroquinones (H_2Q) are much more resistant to electron removal because of their release of a proton via the solvent matrix; the more basic the solvent, the less positive the oxidation potential.

$$H_2Cat \xrightarrow[\text{sol} \to \text{Hsol}^+]{-e^-} [HSQ\cdot] \xrightarrow[\text{sol} \to \text{Hsol}^+]{-e^-} Q \qquad (E_{p,a})_{DMF},\ +0.97\text{ V};\ (E_{p,a})_{AN},\ +1.19\text{ V} \quad (173)$$

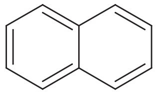
In acetonitrile, the residual H_2O is the stronger base, which results in the formation of H_3^+O.

The electrochemistry of quinones is surprisingly similar to that of dioxygen. It is as if a conjugated carbon link is inserted between two oxygen atoms ($Q \rightarrow SQ^-\cdot \rightarrow HCat^-$ vs $\cdot O_2\cdot \rightarrow O_2^-\cdot \rightarrow HOO^-$; $Q \rightarrow H_2Cat$ vs $\cdot O_2\cdot \rightarrow$ HOOH).

3. Carbonyl Groups, Olefins, and Aromatic Hydrocarbons

Unsaturated carbon centers possess a limited degree of electrophilicity and will accept an electron at potentials significantly more negative than their chloro derivatives (usually at least −2.5 V vs SCE in rigorously anhydrous solvents). When water is present, its reduction is synergistically facilitated via unsaturated carbon. For example, in acetonitrile at a glassy-carbon electrode,

$$H_2O + e^- \longrightarrow [H\cdot] + HO^- \qquad E_{p,c},\ -3.9\text{ V vs SCE} \quad (174)$$



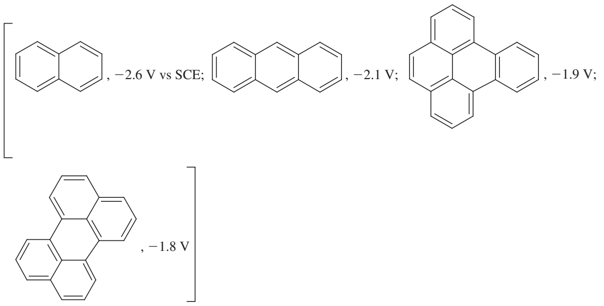
$$+\ e^- \longrightarrow NP^-\cdot \qquad E_{p,c},\ -2.6\text{ V vs SCE} \quad (175)$$

$$NP + H_2O + e^- \xrightarrow{-HO^-} [HNP\cdot] \xrightarrow[H_2O,\ -HO^-]{e^-} H_2NP \qquad E_{p,c},\ \sim -2.0\text{ V vs SCE} \quad (176)$$

Similar synergism is observed for other Brønsted acids (H*A*) in the presence of unsaturated carbon centers.

$$PhCH(O) + HA + e^- \xrightarrow{-A^-} [PhCHOH] \xrightarrow[HA,\ -A^-]{e^-} PhCH_2OH \quad (177)$$

For aromatic hydrocarbons, the more extensive the conjugation is, the lower the LUMO and the less negative the potential for direct electron transfer



4. Brønsted Acids

Carboxylic acids, phenols, and alcohols are electrochemically reduced via their Brønsted acidity at a reduction potential that is a direct measure of their acidity (pK_a) in a given solvent.

$$HA + e^- \longrightarrow 1/2\, H_2 + A^-$$
$$(E_{1/2})_{Pt} \approx -0.059\ (pK_a)\ \text{V vs NHE.} \qquad (178)$$

5. Oxidation of Nucleophilic Substrates and Lewis Bases

All molecules with nonbonding electron pairs (e.g., H_2O, ROH, ROR, RNH$_2$, RSH, RSR, etc.) are, by definition, Lewis bases with a degree of nucleophilicity. Their electrochemical oxidation potential is a measure (1) of the ease of removal for one of the electron pair of electrons and (2) of relative nucleophilicity (the less positive the potential, the more nucleophilic). Aromatic molecules with Lewis-base substituents are easier to oxidize than the aliphatic forms of the substituents (e.g., PhOMe, +1.75 V vs SCE; MeOH, +2.5 V vs SCE) because the aromatic ring provides a means to delocalize the positive charge and electron spin that would result from electron removal (in the case of PhOMe, there are five additional hydrogen atoms to share the positive charge and six unsaturated carbon centers to share the spin density). Within this context, the water molecule in a nonbasic solvent matrix is the most resistant to electron removal,

$$2\,H_2O - e^- \longrightarrow \left[(H_2O)H_2^+O\cdot\right] \longrightarrow HO\cdot + H_3^+O. \qquad (179)$$

Because aliphatic alcohols can be viewed as "organic water" (but with a greater basicity and a weaker O—H bond), they are almost as difficult to oxidize [MeOH(H_2O) → MeO· + H_3^+O + e^-; $E_{p,a}$, +2.5 V vs SCE]. Other aliphatic bases (amines and thiols) are oxidized by similar pathways.

The pi-electron cloud of the aromatic ring is much more susceptable to electron removal than a saturated hydrocarbon. Thus, benzene is oxidized at +2.45 V vs SCE; electron-donating substituents reduce the potential by as much as 0.7 V. Although a rigorously base-free solvent matrix should result in a reversible oxidation of benzene, the resulting cation radical ($PhH^{+}\cdot$) is a strong electrophile and Lewis acid that will react with water (complete separation of water from benzene is difficult; 18 ppm H_2O is a 1-mM concentration) to give PhOH and H_3^+O in a CEC process.

E. Organometallic Molecules

The defining characteristic of organometallic molecules is the presence of one or more metal-carbon bonds. In contrast to the acid/base character of coordination complexes of metal ions (with their ligand-centered redox chemistry, see Section II), the metal-carbon center is highly covalent with limited polarity (similar to carbon-carbon, carbon-nitrogen, or carbon-oxygen centers). As a result, the electrochemistry of organometallic molecules is more closedly related to that of organic molecules than inorganic coordination complexes.

The "foundation stone" of organometallic chemistry is bis(cyclopendienyl) iron(II) [ferrocene, (Cp)Fe^{II}(Cp)], an iron atom sandwiched between two five-membered carbon rings [Cp, $C_5H_5\cdot$; each carbon with a p electron to give (1) two pi-bonds delocalized around the carbon ring and (2) an unpaired electron to give a covalent bond that is shared by the five carbons of the ring]. Thus, the $Fe^{II}(Cp)_2$ molecule has the iron on a line that connects the centers of two parallel planar Cp. groups to give an "iron sandwich."

Figure 7 illustrates the electrochemial redox chemistry in acetronitrile for several coordination complexes of iron [$Fe^{II}(MeCN)_4^{2+}$, $Fe^{III}Cl_3$, and $Fe^{III}(acac)_3$ (acac = acetylacetonate)] in relation to that for two iron organometallics [$Fe^{II}(Cp)_2$ and $Fe^{VIII}(CO)_5$ (iron-pentacarbonyl); both stable 18-electron systems]. In MeCN, $Fe^{II}(MeCN)_4^{2+}$ is the only charged species of the group. It is reversibly oxidized (II/III couple; $E_{1/2}$, +1.6 V vs SCE). The uncharged $Fe^{III}Cl_3$ molecule is reversibly reduced (III/II couple; $E_{1/2}$, +0.2 V vs SCE) to give $Fe^{II}Cl_3^-$, which is reduced by an irreversible two-electron process to iron metal ($E_{p,c}$, −1.5 V vs SCE). The more basic $Fe^{III}(acac)_3$ molecule is reversibly reduced (III/II couple; $E_{1/2}$, −0.7 V vs SCE), but does not exhibit a second reduction peak. The III/II reduction potentials for these three coordination complexes are a measure of their relative electrophilicity (Lewis acidity).

1. Ferrocene

The $Fe^{II}(Cp)_2$ molecule is resistant to reduction, but exhibits a highly reversible one-electron oxidation,

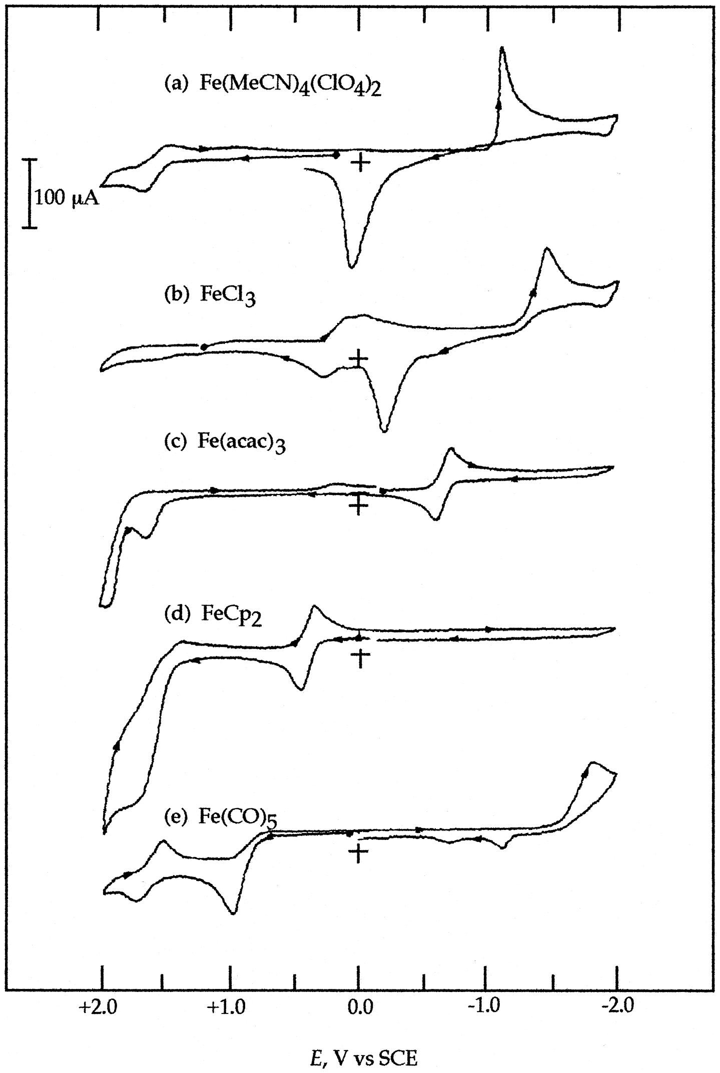


FIGURE 7 Cyclic voltammograms: (a) 3 mM $[Fe^{II}(MeCN)_4](ClO_4)_2$; (b) 3 mM $Fe^{III}Cl_3$; (c) 3 mM $Fe^{III}(acac)_3$; (d) 3 mM $Fe^{II}(Cp)_2$; (e) 3 mM $Fe^{VIII}(CO)_5$ in MeCN (0.1 M tetraethylammonium perchlorate [TEAP]). Conditions: scan rate, 0.1 V s^{-1}; ambient temperature; glassy-carbon working electrode (area, 0.09 cm^2); saturated calomel electrode (SCE) vs NHE, +0.244 V.

$$Fe^{II}(Cp)_2 \xrightarrow[\text{MeCN}]{-e^-} Fe^{III}(Cp)_2^+ \qquad E_{1/2}, +0.45 \text{ V vs SCE}$$
$$S = 0 \qquad S = 1/2 \qquad (180)$$

with the single positive charge delocalized over the 10 equivalent $(Cp)_2$ hydrogens (+0.1 each). For a time there was a belief that the $Fe^{II}(Cp)_2/Fe^{III}(Cp)_2^+$ couple's potential was independent of solvent and, thus, an ideal reference electrode with which to measure solvent effects for other redox couples. However, subsequent measurements have shown that the $Fe^{III}(Cp)_2^+$ ion possesses considerable acidity, which causes some solvent effects. The more serious problem is the limited solubility of $Fe^{II}(Cp)_2$ in H_2O. The respective $E°$ values for the $Fe^{III}(Cp)_2^+/Fe^{II}(Cp)_2$ couple are MeCN, +0.69 V vs NHE (+0.45 V vs SCE); DMF, +0.72 V; py, +0.76 V; Me_2SO, +0.68 V; H_2O, +0.40 V.

2. Iron-Pentacarbonyl

The $Fe(CO)_5$ molecule is equally fundamental to organometallic chemistry and electrochemistry and, like $Fe^{II}(Cp)_2$, is a diamagnetic 18-electron system. It exhibits an irreversible two-electron reduction via residual H_2O (Fig. 7e).

$$Fe(CO)_5 + 2e^- + 2\,H_2O \longrightarrow Fe\,(s) + H_2C(O)$$
$$+\,4\,CO + 2\,HO^- \qquad E_{p,c}, -1.8 \text{ V vs SCE.} \qquad (181)$$

By an anologous process the CO adduct of an iron(II) porphyrin [$(Cl_8TPP)Fe^{IV}(CO)$] is reduced to $H_2C(O)$ at −0.87 V. The oxidation of $Fe(CO)_5$ in MeCN yields $Fe^{II}(MeCN)^{2+}$ in a two-electron process (Fig. 7e),

$$Fe(CO)_5 + 4\,MeCN \xrightarrow{-2e^-} Fe^{II}(MeCN)_4^{2+} + 5\,CO \qquad E_{p,a}, +0.97 \text{ V vs SCE.} \quad (182)$$

On the basis that Fe (s) is oxidized to $Fe^{II}(MeCN)^{4+}$ at ∼0.0 V (Fig. 7a), the carbonyls of $Fe(CO)_5$ stabilize the iron against oxidation by about 22 kcal mol^{-1} [$\Delta E \times 23.06$ kcal mol^{-1} $(eV)^{-1}$; 0.97×23.06]. The $(Cl_8TPP)Fe^{IV}(CO)$ molecule is oxidized at +0.75 V vs +0.32 V for $(Cl_8TPP)Fe^{II}$; a stablization by the CO of about 10 kcal mol^{-1}.

These examples of the electrochemical character of organometallics are limited, but illustrate that their oxidation and reduction is closely similar to that for organic molecules. Thus, the electron transfer is never carbon centered and often involves residual water [H-atom addition via reduction and (HO·) addition or H-atom abstractron via oxidation] or solvent components.

VI. INDUSTRIAL ELECTROSYNTHESIS

A. Electroplating and Electrorefining

1. Copper Refining

Although metallic copper is produced via the smelting of copper ores, it contains enough impurities (and the attendant increase in electrical resistance) to preclude its use as an electrical conductor. Hence, all copper for use in electrical wire and cable must be purified via electrorefining with an impure copper plate as the anode, a pure copper sheet as the cathode, and copper sulfate/sulfuric acid as the electrolyte:

$$\text{Anode: } Cu\ (s)\ (\text{impure}) + 6\,H_2O \longrightarrow Cu^{II}(OH_2)_6^{2+} + 2e^-, \quad (183)$$

$$\text{Cathode: } Cu^{II}(OH_2)_6^{2+} + 2e^- \longrightarrow Cu\ (s)\ (\text{pure}) + 6\,H_2O. \quad (184)$$

Because most copper ores contain traces of gold and silver (which become a part of the impure smelted copper), the sediment that builds up in the bottom of the electrolysis cell concentrates these valuble metals. The periodic refining of the sediment often yields sufficient material to pay for the entire purification process.

B. Aluminum Production

The Hall process for the electrolytic production of aluminum utilizes an iron container that is lined with carbon, which serves as the cathode. The electrolyte within this container is molten cryolite (Na_3AlF_6), which dissolves the aluminum oxide (Al_2O_3) that is the raw material for the process,

$$Na_3AlF_6\ (l) + Al_2O_3\ (s) \longrightarrow 3\,NaAlOF_2. \quad (185)$$

A series of carbon anodes are immersed into the molten solution. Electrolysis produces elemental alumium at the cathode (liquid metal at the operating temperature),

$$\text{Cathode: } Na_3AlF_6\ (l) + 3\,e^- \longrightarrow Al\ (l) + 3\,NaF_2^-, \quad (186)$$

and molecular oxygen at the anode,

$$\text{Anode: } NaAlOF_2 + 2\,NaF_2^- \longrightarrow Na_3AlF_6\ (l) + 1/2\,\cdot O_2 \cdot + 2e^-. \quad (187)$$

Hence, the electrolytic cell reaction only consumes Al_2O_3 to produce Al (l) and $\cdot O_2 \cdot$ (g),

$$\text{Cell: } 2\,Na_3AlF_6 + 3\,NaAlOF_2 \longrightarrow 2\,Al + 3\,Na_3AlF_6 + 3/2\,\cdot O_2 \cdot \quad (Na_3AlF_6 + Al_2O_3)$$

$$\text{Net: } Al_2O_3 \longrightarrow Al + 3/2\,\cdot O_2 \cdot. \quad (188)$$

As the Al_2O_3 is consumed by electrolysis, more of it is continuously added, and liquid aluminum metal is periodically drained from the bottom of the cell.

C. Chloro-Alkali Production

The industrial production of chlorine (Cl_2) and sodium hydroxide (NaOH) involves the electrolysis of molten sodium chloride or of brine solutions. With the molten salt the metallic sodium that is produced at the carbon cathode is vaporized and collected in a condenser as the solid metal. A common system for brine uses a cell with a liquid mercury cathode that flows in one direction along the floor with the brine flowing in the opposite direction past a series of graphite anodes. The electrolysis reactions produce chlorine gas at the anode, which is collected,

$$\text{Anode: } 2\,Cl^-(aq) \longrightarrow Cl_2\ (g) + 2e^-, \quad (189)$$

and sodium amalgam [$Na_2(Hg)$] at the cathode,

$$\text{Cathode: } 2\,Na(OH_2)_6^+ + Hg\ (l) + 2e^- \longrightarrow Na_2(Hg)\ (l) + 6\,H_2O, \quad (190)$$

which flows from the cell and is reacted with water to give a concentrated sodium hydroxide solution,

$$Na_2(Hg)\ (l) + 14\,H_2O \longrightarrow 2\,Na(OH_2)_6^+ + 2\,HO^-\ (aq) + H_2\ (g). \quad (191)$$

D. Organic Molecules and Polymers (e.g., Adiponitrile)

The production of adiponitrile (intermediate in the production of nylon 66) via electrosynthesis is an important industrial process. It also is a good example of how electrogenerated hydrogen atoms [H·] induce the dimerization and oligerimization of terminal olefins. In this process a solution that contains acrylonitrile and a quaternary ammonium salt is circulated through a cathode compartment with a lead electrode. A dilute solution of sulfuric acid is circulated through an anode compartment with a lead-alloy electrode. An ion-exchange membrane separates the two compartments:

$$\text{Cathode: } 4\,\underset{\text{acrylonitrile}}{H_2C{=}CHCN} + 4\,H_2O + 4e^- \rightarrow 2\,\underset{\text{adiponitrile}}{NC(CH_2)_4CN} + 4\,HO^-. \quad (192)$$

$$\text{Anode: } 6\,H_2O \rightarrow 4\,H_3^+O + \cdot O_2\cdot + 4e^-. \quad (193)$$

A portion of the circulating solution from the cathode compartment is continuously withdrawn to an extraction system for removal of adiponitrile.

E. Molecular Hydrogen and Hydrogen Peroxide

Through the use of sulfuric-acid electrolyte solutions, extremely pure molecular hydrogen (H_2) can be electrosynthesized at a platinum cathode. The cathode compartment must be completely free of dioxygen and protected from migration of dioxygen from the anode compartment [Eq. (196)].

$$\text{Cathode: } 4\,H_3^+O + 4e^- \rightarrow 2\,H_2\,(g) + 4\,H_2O \quad (194)$$

If the cathode compartment is saturated with dioxygen (or air) and a graphite cathode is used, then extremely pure solutions of hydrogen peroxide (HOOH) can be electrosynthesized,

$$\text{Cathode: } 4\,H_3^+O + 2\,\cdot O_2\cdot\,(g) + 4e^- \rightarrow 2\,HOOH + 4\,H_2O, \quad (195)$$

$$\text{Anode: } 6\,H_2O \rightarrow 4\,H_3^+O + \cdot O_2\cdot + 4e^-. \quad (196)$$

With both of these processes the ultimate source of hydrogen atoms is water. The anodic reaction replaces the hydronium ions that are removed by the cathodic reactions.

VII. BATTERIES AND FUEL CELLS

Batteries are galvanic cells whereby the chemical energy of the components of the cathode (positive terminal) and the anode (negative terminal) is converted to electrical energy via the cell reaction. Fuel cells are galvanic cells in which the active components of the two electrodes are continuously replenished and the products of the cell reaction are continuously removed.

A. Lead-Acid Battery

The lead-acid battery, which is the electrical-power source for motor vehicles, has exceptional capacity to deliver the large currents necessary to start internal combustion engines. It has a lead dioxide ($Pb^{IV}O_2$) cathode and a metallic lead anode that are immersed in an aqueous sulfuric-acid solution (35% H_2SO_4 by weight).

$$\text{Cathode (+ terminal): } Pb^{IV}O_2\,(s) + 3\,H_3^+O + \underset{[HSO_4^-]}{HOS(O)_2O^-} + 2e^- \rightarrow Pb^{II}SO_4\,(s) + 5\,H_2O \quad (197)$$

$$\text{Anode (}-\text{terminal): } Pb\,(s) + HOS(O)_2O^- + H_2O \rightarrow Pb^{II}SO_4\,(s) + H_3^+O + 2e^- \quad (198)$$

$$\text{Cell: } Pb^{IV}O_2\,(s) + Pb\,(s) + 2\,H_3^+O + 2\,HOS(O)_2O^- \rightarrow 2\,Pb^{II}SO_4\,(s) + 4\,H_2O \quad (199)$$

When current is drawn from the cell via the two electrode reactions, sulfuric acid is consumed in accord with the cell reaction and its concentration decreases. Because the density of H_2SO_4 (l) is almost twice that of water, the state of charge of the battery can be determined by a measurement of the electrolyte density. The initial voltage of a cell is a little greater than 2.0 V. Hence, the conventional 12-V automobile battery has six cells connected in series. A discharged battery is recharged by a generator that causes a current to flow in the opposite direction [Eqs. (197), (198), and (199) each are reversed], which releases H_2SO_4 from the electrodes and increases the density of the electrolyte solution.

B. Edison Battery

The Edison battery is another rechargable high-energy-density system that has superior cycle life relative to the lead-acid battery. Prior to the advent of nuclear power for submarines, large banks of Edison batteries in a series/parallel configuration provided the electric power for underwater propulsion. Relative to the lead-acid battery, the Edison battery has (1) more energy per unit mass (but less energy per unit volume), (2) a smaller cell voltage

(1.3 vs 2.0 V), (3) a constant cell voltage during discharge, and (4) almost unlimited life. In contrast to the sulfuric-acid electrolyte of the lead-acid battery, the Edison battery uses concentrated potassium hydroxide (KOH) as the electroyte. The cathode is nickel dioxide [$Ni^{IV}O_2$ (s)] and the anode is metallic iron. As the cell reaction confirms, the KOH concentration remains constant during discharge, which accounts for the constant cell voltage during discharge (only water is consumed during discharge and restored during recharge).

$$\text{Cathode (+ terminal): } Ni^{IV}O_2\,(s) + 2\,H_2O + 2e^- \rightarrow Ni^{II}(OH)_2\,(s) + 2\,HO^- \quad (200)$$

$$\text{Anode (−terminal): } Fe\,(s) + 2\,HO^- \rightarrow Fe^{II}(OH)_2\,(s) + 2e^- \quad (201)$$

$$\text{Cell: } Ni^{IV}O_2\,(s) + Fe\,(s) + 2\,H_2O \rightarrow Ni^{II}(OH)_2\,(s) + Fe^{II}(OH)_2\,(s) \quad (202)$$

C. "Dry-Cell" Batteries

1. LeClanché Cell

This traditional dry cell consists of a carbon-rod cathode (positive terminal) immersed in a moist paste of $Mn^{IV}O_2$, $Zn^{II}Cl_2$, NH_4Cl, and powdered carbon, which is contained in a metallic zinc-can anode (negative terminal). The voltage (without load) of these cells is about 1.6 V, which have limited shelf life because of corrosion of the zinc can and increased internal resistance.

$$\text{Cathode: } 2\,Mn^{IV}O_2\,(s) + H_2O + 2e^- \rightarrow Mn_2^{III}O_3\,(s) + 2\,HO^- \quad (203)$$

$$\text{Anode: } Zn\,(s) + 2\,NH_4Cl + 2\,HO^- \rightarrow Zn^{II}(NH_3)_2Cl_2 + 2\,H_2O + 2e^- \quad (204)$$

$$\text{Cell: } 2\,Mn^{IV}O_2\,(s) + Zn\,(s) + 2\,NH_4Cl \rightarrow Mn_2^{III}O_3\,(s) + Zn^{II}(NH_3)_2Cl_2 + H_2O \quad (205)$$

2. Alkaline Cell

This modern replacement for the dry cell uses sodium or potassium hydroxide as the electrolyte (in place of the acidic $Zn^{II}Cl_2/NH_4Cl$ electrolyte of the dry cell). Its cathode and anode are essentially the same as the dry cell.

$$\text{Cathode: } 2\,Mn^{IV}O_2\,(s) + H_2O + 2e^- \rightarrow Mn_2^{III}O_3\,(s) + 2\,HO^- \quad (206)$$

$$\text{Anode: } Zn\,(s) + 2\,HO^- \rightarrow Zn^{II}(OH)_2\,(s) + 2e^- \quad (207)$$

$$\text{Cell: } 2\,Mn^{IV}O_2\,(s) + Zn\,(s) + H_2O \rightarrow Mn_2^{III}O_3\,(s) + Zn^{II}(OH)_2\,(s) \quad (208)$$

The alkaline cell has lower internal resistance, longer shelf life, and a larger open-circuit voltage (about 1.9 V) than the dry cell.

D. High-Performance Batteries (Nickel/Cadmium, Silver/Zinc, Lithium)

1. Nickel/Cadmium

This is the most common of the rechargable batteries and has extensive use in "cordless" power tools, electric razors, and electronics. It has similarities to the Edison cell, but better performance in its areas of application.

$$\text{Cathode: } Ni^{IV}O_2\,(s) + 2\,H_2O + 2e^- \rightarrow Ni^{II}(OH)_2\,(s) + 2\,HO^- \quad (209)$$

$$\text{Anode: } Cd\,(s) + 2\,HO^- \rightarrow Cd^{II}(OH)_2\,(s) + 2e^- \quad (210)$$

$$\text{Cell: } Ni^{IV}O_2\,(s) + Cd\,(s) + 2\,H_2O \rightarrow Ni^{II}(OH)_2\,(s) + Cd^{II}(OH)_2\,(s) \quad (211)$$

The main problem with Ni/Cd batteries is their reduced charge capacity if they are not fully discharged before recharging. Because of the toxicity of cadmium, responsible disposal of worn-out batteries is important (but often ignored).

2. Silver/Zinc

Forms of this battery type range from those used in hearing aids, watches, and cameras to those used as portable power sources in space vehicles. The cathode/anode and electrolyte system is similar to that for the alkaline cell, except the $Mn^{IV}O_2$ cathode material is replaced with much more expensive silver oxide (Ag_2^IO). However, it has high stability, a long shelf life, and a large charge density with a stable voltage (about 1.8 V).

$$\text{Cathode: } Ag_2^IO\,(s) + H_2O + 2e^- \rightarrow 2\,Ag\,(s) + 2\,HO^- \quad (212)$$

$$\text{Anode: } Zn\,(s) + 2\,HO^- \rightarrow Zn^{II}(OH)_2\,(s) + 2e^- \quad (213)$$

$$\text{Cell: } Ag_2^IO\,(s) + Zn\,(s) + H_2O \rightarrow 2\,Ag\,(s) + Zn^{II}(OH)_2\,(s) \quad (214)$$

3. Lithium

The development of the lithium/managanese dioxide battery has revolutioned the performance of modern cameras and electronic devices. Because of the use of a dry solvent/electrolyte system (e.g., acetronitrile/LiX), these batteries have operating lifetimes of at least five years, high current capacity, and stable voltages (about 2.5 V) under load and can be used at temperatures as low as −40°C.

$$\text{Cathode: Mn}^{IV}O_2\,(s) + LiX + e^- \rightarrow \text{Mn}^{III}(O)OLi\,(s) + X^- \quad (215)$$

$$\text{Anode: Li}\,(s) + X^- \rightarrow LiX + e^- \quad (216)$$

$$\text{Cell: Mn}^{IV}O_2\,(s) + Li\,(s) \rightarrow \text{Mn}^{III}(O)OLi\,(s) \quad (217)$$

E. Fuel Cells

The hydrogen (H_2)/oxygen ($\cdot O_2\cdot$) anode/cathode combination is the most highly developed fuel cell. It continues to be an essential power source for manned space missions, which accounts for its advanced state of development. Beyond the practical problem of a gaseous fuel (H_2), both electrode reactions require precious-metal catalysts (usually platinum supported on porous carbon electrodes). As indicated in earlier sections, electrochemistry is limited to pathways that involve one electron steps. Hence, the essential function of the electrocatalysts for H_2 oxidation and $\cdot O_2\cdot$ reduction is to provide such pathways for these multi-electron transformations.

$$\text{Cathode: } \cdot O_2\cdot\,(g) + 2\,H_2O + 2\,Pt\,(s) \rightarrow 2\,Pt^{II}(OH)_2\,(s);$$
$$2\,Pt^{II}(OH)_2\,(s) + 4\,H_3^+O + 4e^- \rightarrow 2\,Pt\,(s) + 6\,H_2O \quad (218)$$

$$\text{Anode: } 2\,H_2\,(g) + 4\,Pt\,(s) \rightarrow 4\,PtH\,(s);$$
$$4\,PtH\,(s) + 4\,H_2O \rightarrow 4\,Pt\,(s) + 4\,H_3^+O + 4e^- \quad (219)$$

$$\text{Cell: } \cdot O_2\cdot\,(g) + 2\,H_2\,(g) \rightarrow 2\,H_2O \quad (220)$$

VIII. CORROSION; CATHODIC PROTECTION

All metals (M) react with atmospheric oxygen ($\cdot O_2\cdot$) to form surface films of metal oxides (MO_x). When this film is formed under controlled conditions, it produces an inert (*passivated*) surface that precludes further reaction and corrosion. However, the oxide films on copper alloys and structural steel undergo dissolution when exposed to aqueous media that contain $\cdot O_2\cdot$ and salts, acids, or bases, and the surface no longer is protected and corrodes (dissolves).

A. Structural Steel/Aluminum Anodes

Structual steel in aqueous environments (e.g., oil-production platforms in the Gulf of Mexico) undergoes corrosive dissolution via a number of chemical reactions. The protective oxide coating [Fe_2O_3] is especially susceptible to removal by the acids, salts, and organic matter in seawater, which leaves an exposed iron surface. Two thermodynamically favored reactions are accelerated by the presence of salts and acids to cause rapid corrosion:

$$Fe\,(s) + 2\,H_2O/CO_2 \rightarrow Fe^{II}(OH_2)_6^{2+} + H_2 + 2\,HO(O)O^-, \quad (221)$$

$$2\,Fe\,(s) + \cdot O_2\cdot + 4\,H_2O/CO_2 + 10\,H_2O \rightarrow 2\,Fe^{II}(OH_2)_6^{2+} + 4\,HO(O)O^-. \quad (222)$$

This problem has been solved through the use of a sacraficial metal (usually aluminum) that is attached to the steel structure. The combination constitutes a Galvanic cell with the iron as the cathode (+terminal), the aluminum as the anode (−terminal), and the seawater as the electrolyte. The relevant half reactions and formal potentials [Eqs. (223) and (224)] allow formulation of the cell, its voltage, and the equilibrium constant, K, for the cell reaction [Eqs. (225)–(227)].

$$Fe^{II}(OH_2)_6^{2+} + 2e^- \rightarrow Fe\,(s) + 6\,H_2O \qquad E^{\circ\prime} = -0.44\text{ V vs NHE} \quad (223)$$

$$(H_2O)_5^{2+}Al^{III}OH + CO_2 + 3e^- \rightarrow Al\,(s) + HOC(O)O^- + 5\,H_2O \qquad E^{\circ\prime} = -1.83\text{ V vs HE} \quad (224)$$

$$\text{Cathode (of battery): } Fe^{II}(OH_2)_6^{2+} + 2e^- \rightarrow Fe\,(s) + 6\,H_2O. \quad (225)$$

$$\text{Anode: } Al\,(s) + HOC(O)O^- + 5\,H_2O \rightarrow (H_2O)_5^{2+}Al^{III}OH + CO_2 + 3e^- \quad (226)$$

$$\text{Cell reaction: } 3\,Fe^{II}(OH_2)_6^{2+} + 2\,Al\,(s) + 2\,HOC(O)O^- \rightarrow 3\,Fe\,(s) + 2\,(H_2O)_5^{2+}Al^{III}OH + 8\,H_2O + 2\,H_2O/CO_2\ K \quad (227)$$

$$E^{\circ\prime}_{cell} = \left(E^{\circ\prime}_{Fe(II)/Fe} - E^{\circ\prime}_{Al(III)/Al}\right) = -0.44\text{ V} - (-1.83\text{ V}) = 1.39\text{ V}$$

$$E^{\circ\prime}_{cell} = [0.05915/n]\log K$$

$$\log K = [6/0.05915]1.39 = 141; \quad K = 10^{141}$$

Hence, the steel structure is protected until the aluminum anodes are consumed.

SEE ALSO THE FOLLOWING ARTICLES

ALUMINUM • BATTERIES • CHEMICAL THERMODYNAMICS • CORROSION • ELECTROCHEMICAL ENGINEERING • ELECTROLYTE SOLUTIONS, THERMODYNAMICS • ELECTROLYTE SOLUTIONS, TRANSPORT PROPERTIES •

Electron Transfer Reactions • Fuel Cells, Applications in Stationary Power Systems

BIBLIOGRAPHY

Baizer, M. M., and Lund, H., eds. (1991). "Organic Electrochemistry," 3rd ed., Dekker, New York.

Bard, A. J., and Faulkner, L. R. (2001). "Electrochemical Methods," 2nd ed., Wiley, New York.

Bard, A. J., Parsons, R., and Jordan, J., eds. (1985). "Standard Potentials in Aqueous Solution," Dekker, New York.

Dryhurst, G., Kadish, K. M., Scheller, F., and Renneberg, R. (1982). "Biological Electrochemistry," Vol. 1, Academic Press, New York.

Kissinger, P. T., and Heineman, W. R., eds. (1996). "Laboratory Techniques in Electroanalytical Chemistry," 2nd ed., Dekker, New York.

Kolthoff, I. M., and Lingane, J. J. (1952). "Polarography," 2nd ed., Interscience, New York.

Sawyer, D. T., Sobkowiak, A., and Roberts, J. L., Jr. (1995). "Electrochemistry for Chemists," 2nd ed., Wiley-Interscience, New York.

Weinberg, N. L., ed. (1974, 1982). "Technique of Electroorganic Synthesis," Parts I, II, and III, Wiley-Interscience, New York.

Electron Spin Resonance

Larry Kevan
University of Houston

I. Basic Principles
II. Experimental Aspects
III. Isotropic Hyperfine Analysis
IV. Relation of Hyperfine Constants to Spin Densities
V. Second-Order Hyperfine Effects
VI. Anisotropic Hyperfine Interaction
VII. *g* Anisotropy
VIII. Spin Relaxation
IX. Double-Resonance and Time-Domain ESR
X. Applications
XI. New Developments

GLOSSARY

Dot product Mathematical multiplication operation for two vectors.

ENDOR Abbreviation for the double-resonance technique of electron nuclear double resonance.

ESE Abbreviation for the time-domain electron spin resonance technique of electron spin-echo spectroscopy.

***g* Factor** Constant characterizing the magnitude of the spin angular momentum of an electron or a nucleus.

Hyperfine coupling Interaction energy or coupling between electron and nuclear spins.

Perturbation theory Mathematical approximation method used to simplify the calculation of energy levels from a Hamiltonian operator acting on a wave function for a system.

Spin density Fraction of unpaired electron spin probability at a given nucleus in a molecular system.

Spin Hamiltonian Mathematical operator that operates or acts on spin wave functions of a paramagnetic system to give the energy levels of the system.

Wave function State function that can be used to calculate any property, typically energy levels, of a natural system, typically a molecule.

Waveguide Rectangular pipe of appropriate dimensions to propagate microwaves in a given frequency range.

Zeeman interaction Interaction energy of a magnetic field with the spin angular momentum of a particle.

ELECTRON SPIN RESONANCE is a type of magnetic resonance spectroscopy dealing with transitions between

magnetic energy levels associated with different orientations of an electron spin in an atom or a molecule, generally in an external magnetic field. Measurement of the allowed transitions between the electron magnetic energy levels produces a spectrum of an atomic or molecular system with net electron spin angular momentum. Generally such systems are defined as those having one or more unpaired electrons. Analysis of the electron spin resonance spectrum can give information about the identification of the species, the geometric structure, the electronic structure, and the internal or overall rotational or translational motion of the species. The most common types of systems studied are free radicals, which can be regarded as atoms or molecules containing one unpaired electron, and transition-metal and rare-earth ions. The specificity of electron spin resonance spectroscopy for only species containing unpaired electrons is particularly valuable for the study of chemical reaction intermediates.

I. BASIC PRINCIPLES

In general, the energy of a system or, more specifically, of a molecule depends on momentum. The kinetic energy of a molecule is proportional to the square of the linear momentum, and the rotational energy of the molecule is proportional to the square of the rotational or angular momentum. An isolated electron can be regarded as a point particle that has no classical angular momentum. However, experimentally it is found that isolated electrons in a magnetic field do absorb a quantitized amount of energy, which means that they must have at least two energy levels. These are not translational energy levels, because the amount of energy absorbed does not depend on the kinetic energy of the electron. However, the magnitude of the energy absorbed does depend on the magnitude of the magnetic field to which the electrons are exposed. As an explanation of the existence of these magnetic energy levels, it is postulated that an electron has an intrinsic angular momentum called spin angular momentum. When this spin angular momentum interacts with a magnetic field, two different energy levels are produced whose difference accounts for the absorption of energy by the unpaired electron system.

The Hamiltonian energy operator for the electron spin transition we have just discussed is given by

$$\mathcal{H}_{\text{spin}} = g\beta \mathbf{S} \cdot \mathbf{H}, \qquad (1)$$

where $\mathcal{H}_{\text{spin}}$ is the spin Hamiltonian energy operator; $\mathbf{S}$ is the spin angular momentum; $\mathbf{H}$ is the magnetic field; $g\beta$ is a proportionality constant, where $g = 2.0023$ and is called the g factor or spectroscopic splitting factor, which is dimensionless; and β is the Bohr magneton, which has a value of 9.27×10^{-24} J/T. This spin Hamiltonian operates only on spin wave functions and not on orbital wave functions that are commonly associated with electronic energy levels. For an electron there are two spin wave functions, typically denoted by α and β, which are characterized by $+\frac{1}{2}\hbar$ and $-\frac{1}{2}\hbar$ where $\hbar$ is Planck's constant divided by 2π and gives the unit of spin angular momentum. The two energy levels associated with an electron in a magnetic field are thus given by $+\frac{1}{2}g\beta H$ and $-\frac{1}{2}g\beta H$. The difference between these two energy levels is $g\beta H$, so that the transition energy is given by $h\nu = g\beta H$. This basic electron spin resonance transition is illustrated in Fig. 1.

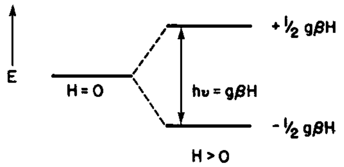


FIGURE 1 Energy-level diagram for ESR.

Typically, electron magnetic resonance is carried out in a magnetic field of about 3000 G (gauss) or 0.3 T (tesla). This corresponds to an energy absorption frequency of about 9 GHz. This frequency is in the microwave range. As for any spectral transition, the number of systems or electrons in the upper and lower energy states at thermal equilibrium is given by a Boltzmann distribution.

II. EXPERIMENTAL ASPECTS

Figure 2 shows a block diagram of a typical electron spin resonance (ESR) spectrometer. It can be seen that this diagram is analogous to that for an optical absorption spectrometer, consisting of a source of electromagnetic optical radiation, a sample cavity, and a detector of the optical radiation intensity. In the block diagram of the ESR spectrometer, the microwave irradiation is produced by a special microwave tube called a klystron or a solid-state device called a Gunn diode. The microwaves are transmitted through a rectangular waveguide or through a special coaxial cable to an isolator, which is a ferrite device allowing microwaves to travel in only one direction. This component is used so as to prevent reflection of the microwaves back into the klystron tube or Gunn diode. The microwaves then pass through a device called an attenuator, which simply controls the intensity of the microwaves and consists of a resistive but electrically conducting element that penetrates to different extents into a waveguide. Although not shown, there also are devices available to shift the phase of the microwaves if that is necessary. The microwaves then pass into a circulator, which is a ferrite device that routes the microwaves in one port and out an

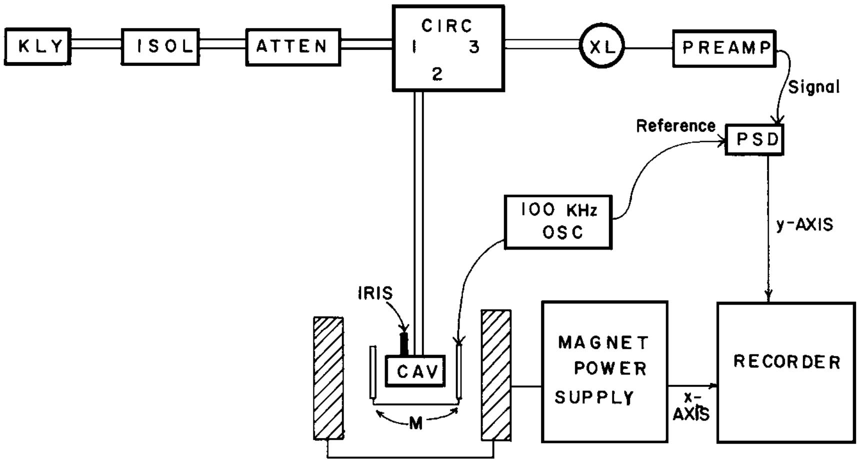


FIGURE 2 Block diagram of a typical ESR spectrometer. KLY, klystron or Gunn diode source of microwaves; ISOL, isolator; ATTEN, attenuator; Circ, three-port circulator; CAV, resonant cavity; M, magnetic field modulation coils; XL, detector crystal; PSD, phase-sensitive detector; and OSC, oscillator. The double lines represent a waveguide connecting the microwave components, and the single line represents a cable connecting the electronic components.

adjacent port in a unidirectional fashion. This component is particularly useful for a resonance-type experiment in which all of the microwaves are absorbed in the sample cavity. When this equilibrium situation is upset by a sample absorbing some of the microwaves in the cavity, an imbalance is set up and some microwaves are reflected from the cavity out through the circulator in a unidirectional fashion to the detector crystal. In the diagram the ESR cavity shown is a reflection type, which is typically more sensitive than a transmission type, which would be more analogous to the sample cell of an optical spectrometer.

The microwaves are detected by a semiconducting crystal diode rectifier mounted in a waveguide. This crystal converts the high-frequency microwave electromagnetic radiation to a direct current voltage, which can then be amplified and treated by ordinary electronics. For optimum sensitivity, the detector crystal is biased with a little reflected microwave power controlled by the size of the iris, which is the coupling hole allowing the microwaves to enter the cavity. The size of this coupling hole can be simply changed with a screw tip.

So that the signal-to-noise ratio can be increased greatly, by approximately a factor of a thousand, magnetic field modulation is used. This is accomplished with small coils mounted on each side of the sample cavity through which radiofrequency radiation, typically 100 kHz, is passed. This produces a modulation of the static magnetic field, which codes the ESR signal at this modulation frequency. Then a device known as a phase-sensitive detector is used, which is referenced to this magnetic field modulation frequency. The microwave signal is coded at this modulation frequency, and the crystal detector converts the signal into a pulsating dc voltage at this modulation frequency. This signal is then passed into the signal channel of the phase-sensitive detector, and this device compares the coding of the signal with that of the reference and passes only signals coded at the same frequency. This has the effect of greatly increasing the signal-to-noise ratio, since noise generated in the system is not coded at the magnetic field modulation frequency. The phase-sensitive detection procedure also has the effect of differentiating the ESR signal and producing a first-derivative curve instead of an absorption curve. The signal from the phase-sensitive detector then goes to a recorder or another readout device and is usually referenced to a swept magnetic field so that one obtains ESR intensity as a function of magnetic field.

III. ISOTROPIC HYPERFINE ANALYSIS

The real power of electron spin resonance spectroscopy for structural studies is due to interaction of the unpaired electron spin with nuclear spins in molecular species. This gives rise to a splitting of the energy levels and generally allows the determination of the atomic or molecular structure of the radical species. In this situation the spin Hamiltonian of Eq. (1) involves additional terms corresponding to the nuclear spin interacting with the magnetic field, to the nuclear spin interacting with the electron spin, and, if the nuclear spin is ≥ 1, to a nuclear quadrupole interaction. This more complete spin Hamiltonian may be written as

$$\mathcal{H}_{\text{spin}} = \beta \mathbf{H} \cdot \mathsf{g} \cdot \mathbf{S} + \sum_n h\mathbf{S} \cdot \mathsf{A}_n \cdot \mathbf{I}_n - \sum_n g_n \beta_n \mathbf{H} \cdot \mathbf{I}_n + h \sum_n \mathbf{I}_n \cdot \mathsf{Q}_n \cdot \mathbf{I}_n. \quad (2)$$

In Eq. (2) the summations are taken over all the nuclei in the molecular species. The new symbols in Eq. (2) are defined as follows: g_n is the nuclear g factor, which is dimensionless; β_n is the nuclear magneton, having units of joules per gauss or per tesla; the nuclear spin angular momentum operator $\mathbf{I}_n$; the electron–nucleus hyperfine tensor A_n; the quadrupole interaction tensor Q_n; and Planck's constant h.

Note that the nuclear Zeeman term involving the interaction of the magnetic field with the nuclear spin angular momentum operator has a negative sign. This is essentially due to the difference in charge between electrons and nuclei. Also note that the nuclear g factor can be either positive or negative, while the electron g factor is taken as intrinsically positive. In Eq. (2) the electron g factor has been written in tensor form involving a 3×3 matrix that connects the magnetic field vector and the electron spin angular momentum vector. Similarly, the hyperfine interaction is written in tensor form connecting the electron spin and nuclear spin angular momentum vectors. The quadrupole interaction is also written in tensor form.

First consider the special case of isotropic hyperfine interaction in which the hyperfine interaction becomes a scalar and can be written in front of the dot product of the nuclear and electron spin angular momentum vectors. For simplicity the electron g factor will also be considered to be isotropic and to be a scalar. This simplification typically applies to most organic and some inorganic free radicals in liquids and also to a few cases in solids. This occurs because rapid tumbling of the molecular species averages out the anisotropic interactions. Also, since the quadrupole interaction is typically small and can only be experimentally resolved in special cases, it will be left out of the simplified spin Hamiltonian. The resulting simplified spin Hamiltonian becomes

$$\mathcal{H}_{\text{spin}} = g\beta \mathbf{H}\cdot\mathbf{S} - \sum_n g_n\beta_n \mathbf{H}\cdot\mathbf{I}_n + \sum_n hA_n \mathbf{S}\cdot\mathbf{I}_n. \quad (3)$$

The external magnetic field is unidirectional, and by convention it is taken to be in the z-direction.

Then the energy levels of the spin system are given by

$$E_n = \int \Psi_n^* \mathcal{H}_{\text{spin}} \Psi_n \, d\tau, \quad (4)$$

where Ψ represents the spin wave functions. The spin wave functions can be taken as products of the electronic and nuclear spin wave functions. These product functions are not exact wave functions for the spin Hamiltonian including the hyperfine interaction term, but they serve as good first approximations. The correct treatment of the problem usually involves some approximation method, and a common one is perturbation theory, which is a standard quantum-mechanical method. Applying first-order perturbation theory to the spin Hamiltonian in Eq. (3) gives the

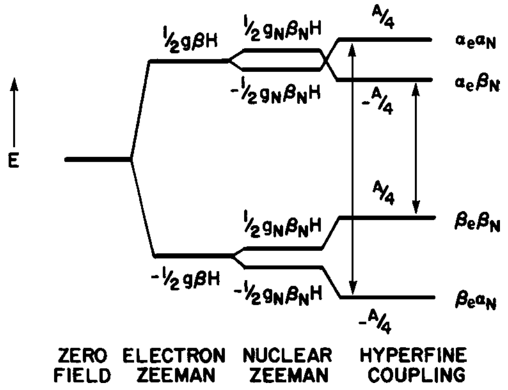


FIGURE 3 Schematic of the first-order spin energy levels of a hydrogen atom, showing successive interactions in the spin Hamiltonian, the allowed ESR transitions, and the spin wave functions.

energy levels given in Eq. (5), where m_S and m_I refer to the electron and nuclear spin angular momentum quantum numbers:

$$E_n = g\beta H m_S - \sum_n g_n\beta_n H m_{In} + m_S \sum_n hA_n m_{In}. \quad (5)$$

The wave functions correct to first-order perturbation theory are just the product functions of the respective electron and nuclear spin combinations. The hyperfine interaction term involves only the z-components of the electron and nuclear spin angular momentum operators when treated by first-order perturbation theory.

This simplified treatment can be applied exactly to a hydrogen atom with $S = \frac{1}{2}$ and $I = \frac{1}{2}$ where the corresponding m_S and m_I values are both $\pm\frac{1}{2}$. The spin wave functions such as $\alpha_e\beta_n$ and the spin energy levels for a hydrogen atom in a magnetic field are shown in Fig. 3.

The transition probabilities between the first-order spin energy levels can be calculated from time-dependent perturbation theory by standard methods in which the magnetic dipole moment operator is used. It is found that the transition moment is only finite for a magnetic dipole moment operator oriented perpendicular to the magnetic field direction. This means that in the design of an electron spin resonance cavity into which the sample is placed, the microwave magnetic field must be arranged to be perpendicular to the external static magnetic field. The second point of interest is that transition moments that determine the selection rules for magnetic resonance are finite for electron spin transitions corresponding to a change in the electron spin orientation but are zero for nuclear spin transitions. These selection rules can be compactly written as

$$\Delta m_S = \pm 1 \quad (6)$$

and

$$\Delta m_I = 0, \quad (7)$$

where m_S and m_I refer to the electron and nuclear spin quantum numbers, respectively.

If we examine Fig. 3 we see that only two of the four possible transitions are allowed, namely, those in which the nuclear spin does not change its orientation. The energy difference between these two transitions is defined as the hyperfine constant, usually symbolized by A in units of megahertz (MHz) or gauss (G). Since $\Delta m_I = 0$, the effect of the nuclear Zeeman term in the spin Hamiltonian will always cancel out for first-order spectral transitions. Thus, this term can be neglected in the Hamiltonian when one is considering only first-order spectra. However, it should be cautioned that if one considers spectra in which the perturbation theory approach must be carried out to second order or if one considers spin relaxation, which will be discussed later, the full spin Hamiltonian must be used.

Figure 4 shows the energy-level diagram based on Eq. (8) for a hypothetical NH^+ species where A^N and A^H are assumed to be positive and $A^N > A^H$. Any subscript refers to the location of the nucleus in the molecule. Figure 4 shows the energy levels of the spin system and the allowed transitions between sets of two energy levels.

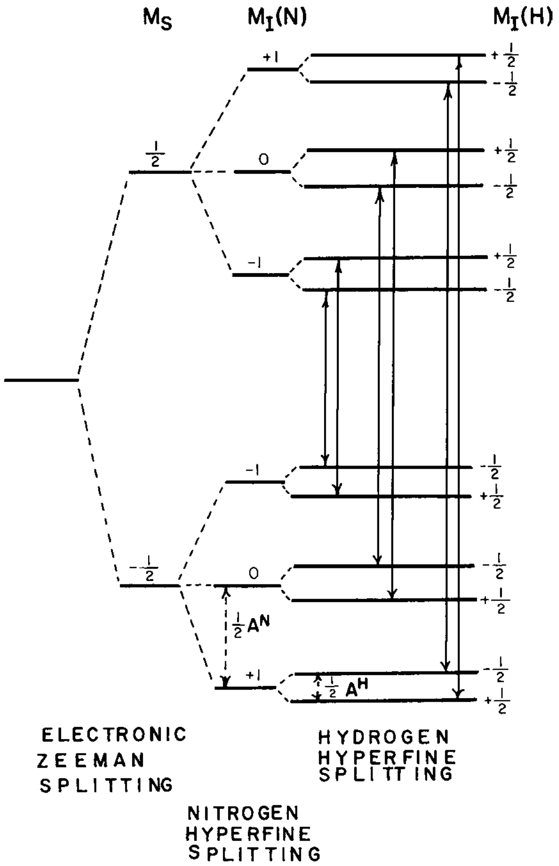


FIGURE 4 First-order hyperfine energy levels and allowed transitions for a hypothetical NH^+ radical with $A^N > A^H > 0$.

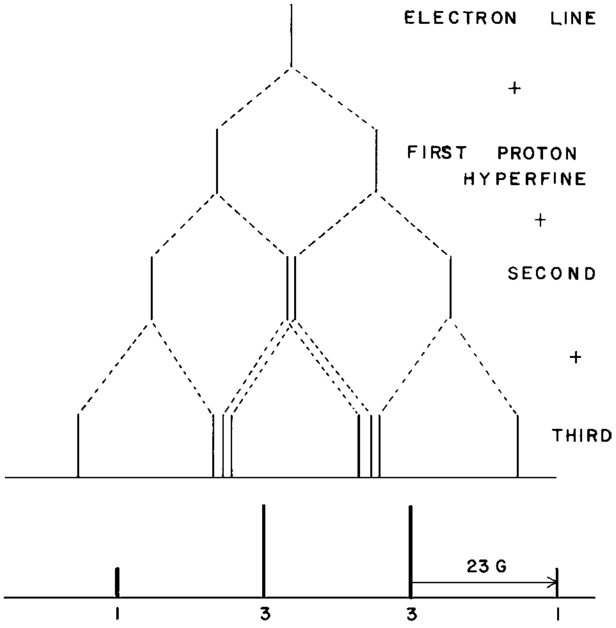


FIGURE 5 Stick diagram for an ESR spectrum of hypothetical NH^+.

So that the actual ESR spectrum can be represented in a simple way, a stick diagram is often used. Figure 5 shows a stick diagram for hypothetical NH^+. Each line represents an observed ESR line. The stick diagram, looks like half of the energy-level diagram, except that the separation between lines in the actual spectrum is equal to the hyperfine constant rather than to some fraction of it. If no lines overlap, the total number of lines is given by $\Pi_n(2I_n + 1)$.

A slightly more complicated situation for hyperfine coupling exists when several nuclei have the same coupling constants. Such nuclei are called equivalent. Consider CH_3 and CD_3 as examples; stick diagrams are shown in Figs. 6

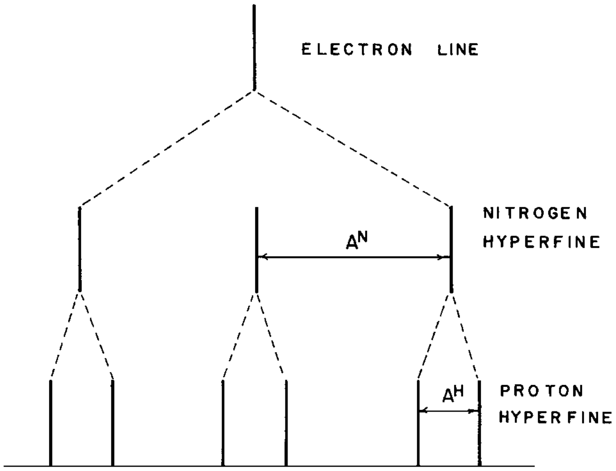


FIGURE 6 Stick diagram for an ESR spectrum of the CH_3 radical.

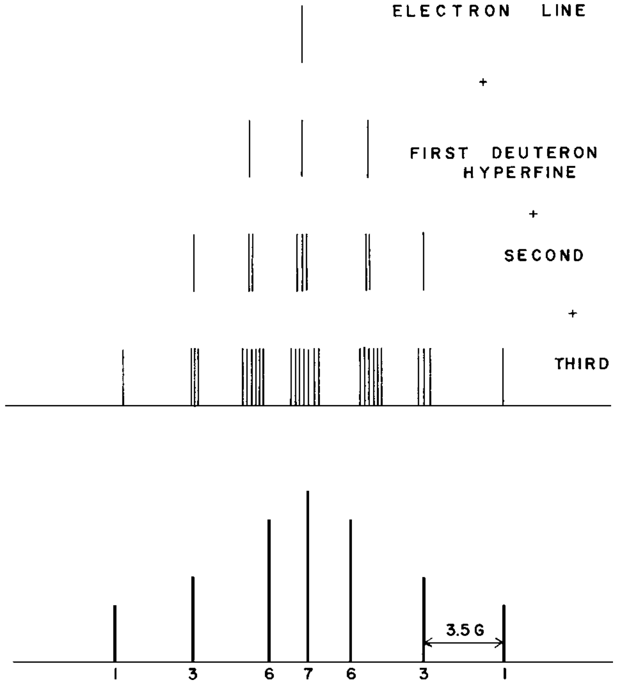


FIGURE 7 Stick diagram for an ESR spectrum of the CD_3 radical.

and 7. After the hyperfine interaction of one nucleus is considered, all the individual lines are no longer separated. The superposition leads to fewer lines and a variable intensity distribution. The splitting between all the lines in the observed spectrum is identical and is equal to the coupling constant. The resulting intensity distributions are characteristic for sets of equivalent nuclei and often simplify analysis. For protons or any nucleus with spin $\frac{1}{2}$, n equivalent nuclei give $n+1$ lines with a binomial intensity distribution. In the general case the number of lines arising from n equivalent nuclei of spin I is given by $2nI+1$.

The equivalence of all magnetic nuclei greatly simplifies the spectrum. For several sets of equivalent nuclei, interpretation is simple only if the difference in coupling constants is large enough to separate the groups. If a radical has two groups of three equivalent protons, the spectrum will be split into four lines by the first group, and each of the four lines will be split again into a quartet by the second group. If the two coupling constants differ enough, the spectrum will appear as four separated quartets. More often the groups (quartets in our example) will overlap. If all the lines are still separated, analysis is tedious but not too complex. Real difficulty arises when the linewidth of the hyperfine lines interferes with their resolution. Unfortunately, this often occurs when the number of interacting protons becomes large.

The assignment of an experimental coupling constant to a given set of nuclei is sometimes not unique. Assignments are usually made on the basis of theoretical calculations or chemical substitutions. The spin density at a particular proton is directly related to its coupling constant. Approximate spin densities can be calculated by a simple Hückel molecular orbital approach. They give a guide to the coupling constant to be expected for a particular position. It is more desirable if the experimental coupling constants can be used to check the accuracy of the theoretical calculation. Hence, assignments based on theoretical spin densities should be used only when no direct information can be obtained. Chemical substitutions can lead to an unambiguous assignment. For example, deuterium can be substituted for a given proton or set of equivalent protons. Deuterium has a spin of 1 and a magnetic moment that is 3.26 times smaller than that of H. Thus deuterium will give a 6.5 times smaller splitting, which is sometimes not even resolved. Methyl groups and Cl can also be substituted for certain protons to delineate the proper coupling assignment.

IV. RELATION OF HYPERFINE CONSTANTS TO SPIN DENSITIES

In spectral analysis, the hyperfine coupling constant A has been treated as an experimental parameter with units of megahertz. The theoretical expression relating the hyperfine coupling constant for atoms is given by

$$hA = \tfrac{8}{3}\pi g\beta g_N\beta_N|\psi(0)|^2, \tag{8}$$

where the units are joules on both sides of the equation and $|\psi(0)|^2$ is the probability density of the electron's being at the nucleus. From a quantum-mechanical point of view, the electron is in contact with the nucleus, and hence the isotropic hyperfine coupling is called a "contact" interaction. Only s orbitals have finite electron density at a nucleus; p, d, and f orbitals have nodes at the nucleus. Thus, contact interaction depends on the s-electron character of the unpaired electron, and the hyperfine constant for a given nucleus provides a measure of the contributions of the s orbitals on the corresponding atom to the total many-electron wave function of the atom or molecule.

For molecules, Eq. (8) can be rewritten as

$$hA = \tfrac{8}{3}\pi g\beta g_N\beta_N\bar{\rho}_N, \tag{9}$$

where $\bar{\rho}_N$ is the unpaired spin density, usually just called spin density, at nucleus N with units of reciprocal volume. The spin density $\bar{\rho}_N$ is evaluated from the total electron wave function and represents the difference between the average number of electrons at the nucleus with α spin ($m_s = +\frac{1}{2}$) and the average number with β spin

($m_s = -\frac{1}{2}$). By convention, the unpaired electron of a free radical is taken to be an α spin; hence spin densities are usually positive. However, a particular nucleus may be in a region of excess β spin, in which case the spin density at that nucleus is negative.

For example, in the benzyl radical the spin density at the meta positions is negative while the spin density at the other ring positions is positive. The sign of the spin density corresponds to the sign of the hyperfine coupling constant. In the typical ESR spectrum, no sign information on the coupling constants is obtained. Nuclear magnetic resonance (NMR) measurements or, under certain conditions, second-order effects in the ESR spectrum are used to determine signs.

By convention, the spin density $\bar{\rho}_N$, which has units of reciprocal volume, is usually normalized by division by $|\psi_N(0)|^2$ to obtain a dimensionless fractional spin density ρ_N, also usually just called spin density. The number ρ_N represents the fraction of unpaired spin on an atom N. A proton hyperfine coupling constant of 142 MHz corresponds to a spin density (ρ_N) of $142/1420 = 0.1$ at the proton where 1420 MHz is $|\psi_N(0)|^2$. The spin density ρ_N may be positive or negative, but $\sum_N \rho_N = 1$ for all spin-$\frac{1}{2}$ radicals.

Since the isotropic hyperfine coupling constant is directly proportional to the s-electron spin density, it can be used to determine orbital hybridization and consequently radical structures. To apply this, one must know what the hyperfine coupling constant is for a 100% s electron on an atom. Only for H atoms is this known exactly. For other atoms, the best available Hartree-Fock wave functions are used to calculate $|\psi_{ns}(0)|^2$. Values have been tabulated that are good to ±10% or much better for the lighter elements.

As an example, consider the trifluoromethyl radical CF_3. To determine the s-electron spin density on the carbon, one must measure the ^{13}C hyperfine coupling constant experimentally. This is found to be 271.6 G, which is ~24% of a full 2s electron on the carbon. This implies near SP^3 bonding in the radical and indicates that CF_3 is tetrahedral and not planar. In contrast, the ^{13}C hyperfine coupling constant for the methyl radical CH_3 is 38 G, which indicates only ~3% s character. This is consistent with a near-planar structure for CH_3. In fact, the time-average structure of CH_3 is planar, but a small amount of s hybridization can arise from out-of-plane vibrations of the H atoms.

Equation (9) shows a direct proportionality between the hyperfine coupling constant and the s-electron spin density. Many radicals have the unpaired electron largely localized in a p orbital, but direct or indirect interaction with orbitals of partial s character can lead to a net spin density at the nuclei. The ethyl radical $\cdot CH_2CH_3$ illustrates two types of mechanisms that lead to hyperfine coupling with all its protons.

The spin polarization mechanism generates observable hyperfine coupling to protons on the carbon containing the p orbital; such protons are called α protons. The two electrons in the C—H_α sigma bond are spin polarized such that the electron nearest the carbon has the same spin as that of the unpaired electron, namely, an α spin. This occurs because the exchange interaction between two parallel spins near the carbon nucleus slightly lowers the energy. This spin polarization causes the spin orientation of the unpaired electron in the p orbital on carbon to be opposite the spin orientation of the bonding electron largely in the hydrogen 1s orbital. Thus the spin density at the proton is negative and the hyperfine coupling constant is negative.

Quantitative calculations for one electron in the $2p_z$ carbon orbital in a $>\dot{C}$—H fragment show that a negative spin density of −0.05 is induced at each α proton. This corresponds to −0.05 (507 G) ≈ −25 G, where 507 G is the value for unit spin density on a proton, which compares well with −23.0 G observed for the methyl radical and −22.4 G observed for the ethyl radical. The negative sign is confirmed by NMR measurements.

In the ethyl radical the β protons on the carbon adjacent to the one with the unpaired electron also produce a hyperfine coupling that is the same order of magnitude as that of the α-proton coupling. A spin polarization mechanism would have to pass through two bonds to reach the β protons and would be expected to be weaker than for α protons. Therefore, an alternative mechanism involving hyperconjugations seems probable. This mechanism can be pictured qualitatively as follows. The unpaired electron is envisioned as occupying a molecular orbital consisting of contributions from the two p_z carbon orbitals. This molecular orbital will overlap H-atom 1s orbitals that are not in the nodal plane of the p_z orbitals and therefore will overlap with the β protons in the ethyl radical. The hyperconjugation mechanism allows some of the unpaired α-spin density to directly overlap into the β-proton 1s orbitals and thus predicts a positive coupling constant. A positive sign is observed by NMR. It should be noted that a positive sign at the β protons is also predicted by the spin polarization mechanism, but the magnitude from this mechanism is expected to be smaller.

In aromatic radicals the unpaired electron is delocalized over the p_z carbon orbitals, so that the spin density in any one p_z orbital is less than it is on a methyl radical. The protons are in the nodal plane of the p_z orbitals and exhibit hyperfine coupling through the spin polarization mechanism. The proton splitting is directly proportional to the p_z or molecular π-orbital spin density on the carbon to which the proton is attached, as represented by Eq. (10):

$$A^H = Q^H \rho_C^\pi, \tag{10}$$

where Q^H is a constant with a value of about −25 G. Equation (10) applies to protons bonded to any atom, except that Q^H is dependent on the atom. It has been most extensively tested for protons bonded to carbon and to a lesser extent for protons bonded to nitrogen.

Although proton hyperfine interactions are by far the most common in aromatic radicals, a great deal of information about spin densities can also be derived from hyperfine interactions with ^{13}C, ^{14}N, ^{19}F, ^{17}O, and so forth. The simple relation of Eq. (10) for protons does not generally apply to these nuclei because the interactions of the unpaired spin with spin density on adjacent nuclei and with lone-pair π electrons must be considered.

V. SECOND-ORDER HYPERFINE EFFECTS

When the hyperfine coupling constants are large and the linewidth is small, second-order hyperfine effects must be considered to explain the observed spectra in many cases. A simple example is the spectrum of a hydrogen atom. The second-order effects in the hydrogen spectrum cause a shift of the center of the spectrum to lower field and cause the observed hyperfine splitting to be larger than the actual value for a field-swept spectrum. Second-order analysis is required because the hyperfine splitting of 507 G is a significant fraction of the typical 3300-G magnetic field used to observe the spectrum. The second-order perturbation theory correction factor to the energy levels is proportional to the square of the hyperfine coupling constant divided by the observing field. This causes a second-order shift in two of the four hydrogen-atom energy levels, as shown in Fig. 8. The two transition energies are increased, so the entire spectrum is shifted to lower field due to this effect. For hydrogen atoms the shift is about 18 G, which corresponds to a change in the apparent g factor of 0.0108.

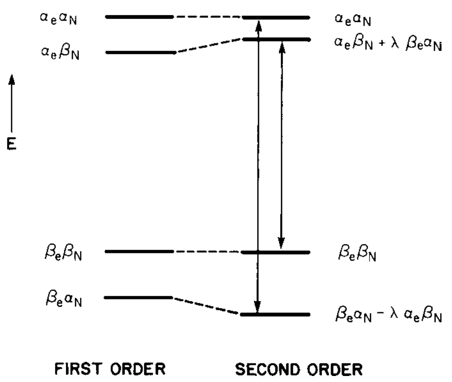


FIGURE 8 Schematic of the first- and second-order energy levels of a hydrogen atom, the first-order allowed ESR transitions, and the first- and second-order spin wave functions. The magnitude of the second-order shift is $A^2[4(g\beta H + g_N\beta_N H)]^{-1}$.

The difference between the allowed transitions shown in Fig. 8 is not changed from the first-order situation at constant magnetic field. However, in an actual ESR spectrometer the magnetic field is swept, and since the second-order correction term depends on the magnetic field, the "apparent" splitting will change in second order. This observed apparent splitting will be larger than the actual hyperfine splitting, and for the case of hydrogen atoms this difference is about 2 G.

The second-order effect on the hydrogen-atom spectrum is a subtle one, in the sense that the nominal appearance of the spectrum does not change. Generally in more complex paramagnetic systems additional lines in the ESR spectrum appear that are due to second-order effects in the energy-level diagram. These types of transitions must be identified in order to interpret a spectrum well enough to assign the structure of a radical. The "extra" lines that commonly occur due to second-order hyperfine effects can be divided into two classes. One class arises from the splitting of some of the degeneracies of the inner lines of a hyperfine pattern involving several equivalent nuclei. The second class involves forbidden transitions where the nuclear spin selection rule is violated and transitions corresponding to $\Delta m_I = \pm 1$ are observed.

In general, extra spectral lines due to the splitting of the degeneracies of equivalent nuclei occur when the second-order hyperfine correction magnitude is greater than the linewidth. This has been observed only in liquid-phase spectra. To second order, the transition energies for a system with one unpaired electron and one type of degenerate nuclei of nuclear spin I is given by

$$E_n = g\beta H + AM_I + \frac{1}{2}\frac{A^2}{g\beta H}\left[I(I+1) - M_I^2\right], \qquad (11)$$

where M_I is the total nuclear spin quantum number for the set of equivalent nuclei. For n equivalent nuclei in general $I^2 \neq M_I^2$, so from Eq. (11) one can see that some splitting of the first-order degeneracy will result. An example of this situation occurs for the trifluoromethyl radical in solution, where the first-order spectrum would predict four equally spaced lines with relative intensities of 1 : 3 : 3 : 1. Since A_{iso} for ^{19}F is 145 G, which is a large value, second-order effects are observable. In the actual spectrum at sufficiently high resolution six lines are observed with relative intensities 1 : (1 : 2) : (1 : 2) : 1, where the (1 : 2) lines are two closely spaced lines whose splitting is due to second-order shifts. All lines are shifted slightly to low field, but different energy levels are shifted by different amounts, which leads to the additional splitting of about 9 G in this case.

The other result of second-order hyperfine effects is to mix the first-order wave functions as shown in Fig. 8 so as to partially allow some forbidden transitions that involve flipping of nuclear spins. When the first-order wave

functions are mixed, the normally forbidden transitions become weakly allowed, due to components of the oscillating microwave magnetic field parallel to the direction of the external magnetic field, instead of perpendicular to the external magnetic field as is the case for ordinary allowed ESR transitions. In the practical situation there is usually a small component of the applied microwave magnetic field in the direction parallel to the external magnetic field, which can bring about such nominally forbidden transitions. These types of transitions are seldom seen in liquids but are relatively common in solids.

VI. ANISOTROPIC HYPERFINE INTERACTION

The general spin Hamiltonian was given by Eq. (2), in which the interaction parameters were written in the general tensor form. The total hyperfine tensor A can be represented by a 3×3 matrix that connects the three components each of the electron spin angular momentum and the nuclear spin angular momentum. The hyperfine tensor is a real matrix and can always be diagonalized. Thus a general tensor referenced to an x–y–z axis system can be written as

$$\mathsf{A} \equiv \begin{vmatrix} A_{xx} & A_{xy} & A_{xz} \\ A_{yx} & A_{yy} & A_{yz} \\ A_{zx} & A_{zy} & A_{zz} \end{vmatrix}. \qquad (12)$$

This general tensor can be transformed to another axis system in which the new tensor is diagonal. This requires finding the proper transformation matrix L, which diagonalizes the A tensor as

$$\mathsf{L} \cdot \mathsf{A}_{xyz} \cdot \mathsf{L}^{\mathrm{T}} = {}^{\mathrm{d}}\mathsf{A}_{\alpha\beta\gamma}, \qquad (13)$$

where L^{T} is the transpose of L. The diagonalized tensor ${}^{\mathrm{d}}\mathsf{A}$ is diagonal in the α–β–γ axis system, and the components of L are the three sets of direction cosines that relate the x–y–z axis system to the α–β–γ axis system.

The components of the diagonalized hyperfine tensor consist of an isotropic part A_0 and a purely anisotropic part A', whose orientational average is zero. This decomposition is shown in Eq. (14):

$$\begin{aligned} {}^{\mathrm{d}}\mathsf{A} &= \begin{vmatrix} A_{\alpha\alpha} & 0 & 0 \\ 0 & A_{\beta\beta} & 0 \\ 0 & 0 & A_{\gamma\gamma} \end{vmatrix} = \begin{vmatrix} A_0 & 0 & 0 \\ 0 & A_0 & 0 \\ 0 & 0 & A_0 \end{vmatrix} \\ &= \begin{vmatrix} A'_{\alpha\alpha} & 0 & 0 \\ 0 & A'_{\beta\beta} & 0 \\ 0 & 0 & A'_{\gamma\gamma} \end{vmatrix}. \end{aligned} \qquad (14)$$

Thus the sum of the diagonal elements of ${}^{\mathrm{d}}\mathsf{A}$ gives $3A_0$, and the sum of the diagonal elements of A' is zero. The diagonal elements of a diagonalized hyperfine tensor are called the principal values.

To determine the anisotropic hyperfine tensor experimentally, one must usually use a single crystal, although if there is only one interacting nucleus or possibly more with sufficiently large anisotropy, it is sometimes possible to determine the tensor from powder spectra. For single crystals it is necessary to measure the angularly dependent hyperfine splitting in three mutually perpendicular planes with respect to the external magnetic field. From this data there are well-known procedures to obtain the hyperfine tensor in the axis system chosen for measurement. Then, as outlined above, this tensor may be diagonalized and the principal values with their associated direction cosines may be determined.

The physical interpretation of the anisotropic hyperfine principal values is given by the classical magnetic dipolar interaction between the electron and nuclear spin angular momenta. This interaction energy is given by

$$\mathscr{H}_{\mathrm{aniso}} = -g\beta g_N \beta_N \left[\frac{1 - 3(\cos^2 \phi)}{r^3} \right] I \cdot S, \qquad (15)$$

where r is the vector between the unpaired electron and the nucleus with which the interaction occurs and ϕ is the angle between r and the electron spin angular momentum vector S, which is in the direction of the external magnetic field. The A' principal values are given by

$$A' = -g\beta g_N \beta_N h^{-1} \left\langle \frac{1 - 3(\cos^2 \phi)}{r^3} \right\rangle_{\mathrm{av}}, \qquad (16)$$

where av denotes a spatial average over the electronic orbital of the unpaired electron. The three components of A' are given by three different values of $\cos \phi$ corresponding to rotation in three mutually perpendicular planes of the principal axis system.

The dipolar function $\langle 1 - 3(\cos^2 \phi)/r^3 \rangle\rangle_{\mathrm{av}}$ can be evaluated from known wave functions of electrons in s, p, d, and other orbitals on different atoms. For s orbitals the dipolar function is zero because of spherical symmetry. The cylindrical symmetry of p orbitals gives three components, $A'_{\parallel}$, $A'_{\perp}$, and $A'_{\perp}$, which are related by $A'_{\parallel} = -2A'_{\perp}$. Note that the dipolar angular function changes sign at $\phi = 54^\circ 44''$. Thus, the space around a nucleus can be divided into four regions alternating in sign. In the determination of the hyperfine tensor, a set of signs of the components will be obtained so that the sum of the diagonal principal values is zero.

The A' components from Eq. (16) have been evaluated theoretically for unit spin density in atomic radial wave functions for p orbitals. Thus the experimental anisotropic hyperfine components can be used to estimate the amount of spin density in p orbitals. This complements the use

of isotropic hyperfine values to obtain the amount of spin density in s orbitals.

VII. *g* ANISOTROPY

In the general spin Hamiltonian given by Eq. (2), the g factor given in the electron Zeeman energy term is written as a tensor connecting the electron spin angular momentum operator **S** and the magnetic field vector **H**. A free electron has *only* spin angular momentum, and its orientation in a magnetic field is determined only by this physical property. However, in general, in atomic and molecular systems there will be some contribution from orbital angular momentum to the total unpaired electron wave function. In this case the orbital and spin angular momentum vectors interact, and by convention this interaction is incorporated into an "effective" anisotropy in the g factor. In this representation the spin angular momentum vector **S** no longer represents "the true spin" because the true spin has only spin angular momentum and is associated with an isotropic g value. Instead, when g is written as a tensor the spin angular momentum vector represents an effective spin, which instead of being oriented along the magnetic field direction is oriented along the vector $\mathbf{H} \cdot \mathsf{g}$. For most purposes this nuance will not affect our utilization of the g tensor formulation.

The experimental determination of the g tensor is carried out by a procedure completely analogous to that for determination of the anisotropic hyperfine tensor. Measurements are required as a function of angle in three mutually perpendicular planes. From this data, a general g tensor is obtained, which is diagonalized to find the principal values. The principal axes of the g tensor are often the same as for the hyperfine tensor, but they do not have to be.

The interpretation of the principal value of the g tensor can be conveniently discussed by Eq. (17):

$$g_{\mathrm{obs}} = g_e + \frac{C\lambda}{\Delta E}. \tag{17}$$

In this expression, g_e is the g factor for an isolated spin (2.0023), λ is the spin–orbit coupling constant, C is a proportionality constant calculated from the electronic wave functions, and ΔE is the energy difference between the ground state and the first excited state. Values of λ have been obtained for a number of atoms and ions from atomic spectra, but the particular value to be used in a molecular system can only be approximated by this. In general, λ values increase with atomic number. The values of ΔE can sometimes be deduced from electronic spectra. Thus the g anisotropy is related to the electronic wave function, and if sufficient information is known about the electronic wave function the principal g components can be calculated and compared with experimental data. The difficulty is that information about the excited-state energy levels needs to be known to properly calculate the g tensor and this is generally known only for simple molecular systems. In a few cases, such as for the CO_2^- radical ion, detailed calculations have been carried out and the experimental g anisotropy has given information about the molecular wave function.

Organic radicals generally have the unpaired electron in a p orbital, which has orbital angular momentum. However, the weak molecular electrostatic field splits the M_L components and gives $M_L = 0$ as the lowest state. In this case $g \simeq g_e$. Nevertheless, small deviations from g_e do occur and can be readily measured, particularly in single crystals, since at a typical 3300-G field, $g = 0.0006$ for a 1-G shift.

When an organic radical contains an atom with a large spin–orbit coupling constant, such as oxygen, sulfur, or halogens, the g anisotropy becomes significantly greater and the average g value generally shifts to larger values. This serves as a diagnostic tool for radicals in solution and in solids. One particular example involves peroxy radicals in which the unpaired spin is localized largely on the oxygen and for which the average g equals 2.015 in a wide variety of environments. Sulfur-containing radicals also often show large g anisotropy. This can be used as a diagnostic test for the localization of the radical site in some biological molecules that contain sulfur.

The largest g anisotropy occurs for transition-metal ions, where the g anisotropy is very useful for discriminating between transition-metal ions in different types of environments. The range of g anisotropy can be rather large. Typical values for axial g anisotropy range from $g_\perp = 2.04$ and $g_\parallel = 2.17$ for copper complexes to $g_\perp = 6$ and $g_\parallel = 2$ for some ferric complexes.

VIII. SPIN RELAXATION

The energy between the magnetic energy levels at 3000 G, $g\beta H$, is only 10^{-3} of kT at 300 K. At thermal equilibrium the Boltzmann factor, $\exp(-g\beta H/kt)$, gives the population ratio of the two levels, so the levels are almost equally populated. The application of microwave energy causes transitions between the magnetic levels. The microwave field stimulates transitions in both directions with a probability that depends on the microwave power and on the number of spins in each level. Transitions from the lower to upper levels absorb energy, while upper- to lower-level transitions emit energy. Since the population is slightly greater in the lower level there will be a net absorption of microwave energy; this provides the observed ESR signal. Under steady application of the microwave field with no

other interactions, the populations in the magnetic energy levels would soon become equal; there would then be no net absorption of microwave energy and no ESR signal.

However, the spin system is subject to other interactions, the very interactions that bring about thermal equilibrium. These interactions can be collectively called spin–lattice interactions. They comprise radiationless interactions between the spin system and the thermal motion of the "lattice" or surroundings. The inverse of the rate of spin–lattice induced transitions is described by a characteristic time called the spin–lattice relaxation time and is denoted by the symbol T_1.

At sufficiently low microwave powers, the spin–lattice relaxation processes are fast enough to maintain a thermal equilibrium population between magnetic energy levels. As the microwave power is increased the net upward rate of microwave-induced spin transitions from the lower to upper states is increased and eventually competes with the spin–lattice induced net downward rate. The spin populations in the two magnetic states become more equal and the ESR signal intensity decreases; this is known as power saturation. Normally, one wants to use low enough microwave power to avoid power saturation.

In addition to spin–lattice relaxation, in which energy is transferred from the spin system to the lattice, there exist spin–spin relaxation mechanisms, in which energy is redistributed within the spin system. One may think of this redistribution as a modulation of the spin energy levels. In both fluid and solid phases, the net local magnetic fields are rapidly varying due to different types of molecular motion, and a given spin level at $m_S g\beta H$ is therefore modulated. At high spin concentrations, direct spin–spin exchange and dipolar interaction can also occur. The characteristic time for spin–spin relaxation within a single spin system is symbolized by T_2.

In a single spin system the spin–lattice (T_1) and spin–spin (T_2) relaxation times can be given a precise classical and quantum-mechanical description. A collection of spins has a magnetic moment vector $\mathbf{M}$, which can be resolved into three components, M_x, M_y, and M_z. Before a magnetic field is applied, the number of spins in the two magnetic energy states is equal; after the field is applied, some of the spins begin flipping to achieve a thermal equilibrium distribution between the two states. For an applied magnetic field in the z-direction the spin flips cause M_z to change toward a steady value M_0, which is proportional to the measured static magnetic susceptibility. M_z approaches M_0 with a time constant T_1 such that $M_z = e^{-1}M_0 = 63\% M_0$ in time T_1. So that resonance can be observed, the microwave magnetic field H_1 is applied perpendicular to H_z. If the intensity of H_1 is increased greatly with a pulse of microwaves, the spin system saturates. This means the populations in the upper and lower spin states are equalized, $M_z = 0$, and the resonance absorption disappears. After the pulse, the recovery of M_z toward M_0 with a time constant T_1 can be observed by the growth of the resonance line. The term T_1 is also called the longitudinal relaxation time, because it refers to relaxation along the magnetic-field axis.

The M_x and M_y components of $\mathbf{M}$ are not changed by a spin flip. The m_x and m_y components of each individual spin are randomly oriented before and after the magnetic field H_z is applied. However, application of H_1 in the x–y plane can produce a net phase alignment of the m_x and m_y components to give M_x and M_y. When H_1 is removed, the phase coherence of the spins decays by 63% in time T_2. The term T_2 is also called the transverse relaxation time because it refers to relaxation of magnetization components transverse to the external magnetic field.

An ESR line is not infinitely sharp; it has a shape and width due to spin relaxation. The equations of motion for M_x, M_y, and M_z in the presence of an applied field H_0 and including the spin relaxation processes discussed above are called the Bloch equations. The solution to these equations predicts a Lorentzian line with a halfwidth at halfheight of T_2^{-1}. Lorentzian lineshapes are indeed often found for free radicals in liquids. In this case T_2 can be determined from the linewidth. The Bloch equations also predict how the ESR signal intensity will vary with increasing microwave power. The ESR signal increases, reaches a maximum, and then decreases with increasing microwave power; this behavior is called power saturation. From an analysis of the power saturation curve of ESR intensity versus microwave power, it is possible to determine T_1.

In solids, typical ESR lineshapes are Gaussian instead of Lorentzian. One common interpretation of the Gaussian lineshape is that it is composed of a distribution of Lorentzian lineshapes, each of which corresponds to a group of spins forming a "spin packet" which "see" the same local magnetic environment. If these spin packets are randomly distributed in intensity they will superimpose to give a Gaussian lineshape. Note that for Gaussian lines T_2 cannot be determined from the linewidth. Gaussian lines still undergo microwave power saturation, but very careful and sometimes complex analysis is required to extract values of T_1 and T_2.

A more direct method to obtain values of the spin–lattice and spin–spin relaxation times is to use time-domain ESR methods, which are briefly described next.

IX. DOUBLE-RESONANCE AND TIME-DOMAIN ESR

Double-resonance experiments are usually carried out in spectroscopy to increase spectral resolution. There are

several different kinds of double-resonance experiments that can be done in ESR. The technique that appears to be the most useful currently and one that is becoming more widespread is electron nuclear double resonance, abbreviated ENDOR. Essentially, ENDOR is the observation of the effect of applying a second frequency, which induces nuclear spin flips, simultaneously with the microwave frequency that induces electron spin flips. The observation of ENDOR depends on partial saturation of an ESR transition. In simple terms, the ENDOR effect is simply the change of the ESR intensity when a second radio frequency field is applied to the system.

Experimentally, ENDOR is carried out by fixing the external magnetic field at an ESR hyperfine line and then sweeping the radio frequency from perhaps 1 to 20 MHz. ENDOR lines will be observed corresponding to differently coupled nuclei in the paramagnetic system. For weak hyperfine interactions, the ENDOR lines for each type of spin-$\frac{1}{2}$ nucleus are given by $\nu_N \pm A/2$ where $\nu_N = g_N \beta_N H/h$. For protons, $\nu_\mathrm{H} \approx 14$ MHz at 3300 G. Thus a pair of lines is observed in the ENDOR spectrum for each type of coupled protons. The hyperfine constant A can be measured more accurately with ENDOR than by ESR, because one is measuring megahertz instead of gigahertz. Also, the hyperfine nucleus can be identified from the value of g_N, since these values are characteristic for different nuclei.

The practical advantages of ENDOR experiments can be summarized as follows:

1. Increased resolution can be obtained, which is particularly important for inhomogeneously broadened lines typically observed in solids where the hyperfine structure is unresolved. A typical ESR linewidth in solids is 10 G, whereas a typical ENDOR linewidth is about 0.03 G. A classic example is that of a trapped electron or of trapped hydrogen atoms in potassium chloride crystals. In these paramagnetic systems, ENDOR couplings to different shells of magnetic nuclei are resolved, and a detailed picture of the electron and hydrogen-atom wave functions can be obtained. By ordinary ESR, only a broad line is observed, in which the hyperfine information for the different magnetic nuclei in the solid is not resolved.

2. Hyperfine constants can be measured more accurately by ENDOR than by ESR. This is particularly important for measuring very small hyperfine couplings. Also, very small changes in the hyperfine couplings due to temperature and so forth are most accurately measured by ENDOR.

3. ENDOR can lead to significant spectral simplification, which is particularly important in looking at radicals in liquids. This is because the total number of spectral lines is much less in ENDOR than in ESR. In ENDOR the total number of lines is additive for different types of coupled nuclei, whereas in ESR the number of lines is multiplicative. A specific example can be considered for the triphenylmethyl radical. In the ESR spectrum the total number of lines is 196, spread over a 30-G range. By ENDOR there are only 6 lines, which can be easily resolved because you see only one pair of ENDOR lines for each group of equivalent nuclei. In the triphenylmethyl radical there are three different types of protons—namely, meta, ortho, and para protons—each of which have different hyperfine couplings. Thus this spectral simplification leads to effectively increased resolution for liquid-phase spectra.

4. ENDOR is also very useful for studying details of spin relaxation mechanisms. The effect of molecular motions on relaxation processes can be studied. Solvation and temperature effects on both electron and nuclear relaxation mechanisms can also be investigated.

Time-domain ESR methods have become quite important for modern applications of ESR. In a time-domain experiment one uses microwave pulses rather than steady-state microwaves. It is possible to use pulses as short as a few nanoseconds, which makes fast kinetic processes involving paramagnetic species observable and allows the direct measurement of relaxation times. In addition, time-domain ESR methods have been useful for more direct determination and study of relaxation mechanisms and for developing new methods for obtaining structural information in disordered systems.

Two methods of time-domain ESR will be briefly discussed. The oldest method is that of saturation recovery, which is a direct method to determine spin–lattice relaxation times. The idea is to perturb the steady-state population of spins with a partially saturating pulse of microwaves and then to observe with a very weak microwave field the recovery of the perturbed spin population to equilibrium. In the absence of complications, the recovery process is exponential and can be related to the time constant for spin–lattice relaxation. Exponential recoveries are generally observed in liquids and in some cases in solids. If the spin–lattice relaxation time is not much longer than the spin–spin relaxation time, which is atypical in paramagnetic systems, the interpretation of saturation recovery data becomes more complex.

The second type of time-domain method that will be discussed is electron spin-echo spectroscopy, abbreviated ESE. Figure 9 gives an illustration of a two-pulse electron spin-echo response. Two resonant microwave pulses are applied to the system with typical pulse lengths of 10–100 nsec and pulse powers corresponding to several hundred watts. The first pulse (1) essentially starts a clock and flips the spins from the external magnetic field direction

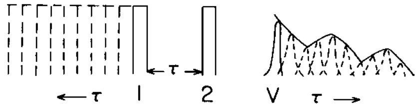


FIGURE 9 Illustration of a two-pulse electron spin-echo signal. Microwave pulses 1 and 2 separated by time τ produce the echo signal V at a second time τ after pulse 2. As τ is increased the echo amplitude generally decreases, and in solids the amplitude may be modulated as shown here. The decrease in echo amplitude is related to transverse magnetic relaxation times, and the modulation is related to weak anisotropic hyperfine coupling to nearby magnetic nuclei.

along the z axis to the x–y plane. During time τ between the pulses the spins precess in the x–y plane. After time τ a second pulse (2) is applied, which is a 180° pulse and which flips the spins into the other direction in the x–y plane. The spins then precess back together, and at a second time τ after the second pulse they coalesce and form a burst of microwave energy called an echo (V). As the time between the pulses is increased, the echo intensity decreases; in liquids this decrease is exponential with a time constant that gives the spin–spin relaxation time T_2. In solids the decay behavior is usually more complex and is only indirectly related to T_2.

In solids the decrease in echo intensity is often modulated with increasing τ, as shown in Fig. 9; this modulation is related to weak anisotropic hyperfine interactions with surrounding nuclear spins. It is of particular interest that this modulation is retained in a disordered sample such as a powder or a frozen solution, so that this technique provides an interesting new approach to obtain structural information about paramagnetic species in disordered systems. The interpretation of electron spin-echo modulation patterns has been used to determine detailed geometric information about the solvation structure of paramagnetic species such as metal cations, molecular anions, and even solvated electrons. This electron spin-echo modulation technique has also been used to study the coordination of paramagnetic species on catalytic oxide surfaces, as well as in a variety of other systems of practical interest. The information from electron spin-echo modulation analysis is essentially the same as what one would obtain from resolved ENDOR spectra in disordered systems. However, in most actual disordered systems, resolved ENDOR is usually not seen, which demonstrates the advantage of the electron spin-echo modulation method.

In addition to the two-pulse spin echo illustrated in Fig. 9, more-complicated pulse sequences are now being routinely used. In a three-pulse sequence, the second 180° pulse in a two-pulse sequence is split into two 90° pulses separated by time T. Then the first experimentally controllable interpulse time τ can be adjusted so as to eliminate one nuclear modulation while the second interpulse time T is swept to produce the echo modulation pattern. This can simplify analysis of the pattern. Also, the echo decay becomes slower since it is usually dominated by the spin–lattice relaxation time T_1. This results in more detectable modulation periods and hence more accurate analysis. A four-pulse sequence in which an additional 180° pulse is introduced in the middle of the T period for a three-pulse sequence is also useful for measuring combination frequencies.

Pulsed ENDOR has also become a more common technique. This is achieved by adding a radio frequency pulse within a spin-echo pulse sequence. Then, by detecting the echo intensity while the radio frequency is swept, one can obtain a pulsed ENDOR spectrum which directly reveals electron–nuclear hyperfine frequencies.

X. APPLICATIONS

Electron spin resonance is widely applicable to organic, inorganic, and biological systems. The most common application is probably the identification of paramagnetic reaction intermediates in chemical reactions or in materials after various physical or chemical treatments. This identification is generally possible by determination of the geometric structure of the paramagnetic species by virtue of hyperfine interaction with magnetic nuclei in the paramagnetic species. It is also often desired to determine the location of a paramagnetic species in a solid material. This can be accomplished, in principle, by detecting very weak hyperfine interactions with nuclei in the material and may require double-resonance or time-domain electron magnetic resonance techniques. The location of paramagnetic species in solid systems is particularly important for catalysts, polymers, and frozen systems of biological interest.

Another important application of electron spin resonance is to directly determine the electronic structure of free radicals by measuring spin densities at various locations within the radical species. Experimental spin densities are also used to directly test the validity of approximate molecular wave functions. Electron spin resonance has been one of the major ways to evaluate various quantum-mechanical approximations for the determination of molecular wave functions.

Since electron spin resonance is an excellent analytical method for paramagnetic species and free radicals, it can be used to obtain a variety of kinetic and thermodynamic data. In this respect it is used in the same way as any other spectroscopic technique. Kinetic data can be obtained by studying radical intensity versus time. By using time-domain electron magnetic resonance techniques such as electron spin-echo spectroscopy, one can detect transient species with lifetimes as short as 100 nsec.

Steady-state electron spin resonance can be used to study kinetics over periods of milliseconds to hours. Thermodynamic information can also be obtained by measuring reaction intermediates as a function of temperature.

Another major area of application involves magnetic energy transfer. By measuring magnetic relaxation times of paramagnetic species, one can determine the degree of interaction of paramagnetic species or free radicals with the "lattice" in which they are embedded. This lattice may be a liquid or a solid. The mechanism of the magnetic energy transfer is an important question and has been addressed in many studies. Applications of this type relate to the diffusion of paramagnetic species in liquids and in solids and to the coupling of the spin system to the electronic system of the lattice.

One simple application of magnetic energy transfer that is often neglected is to use microwave power saturation to distinguish overlapping radicals. Radicals of different chemical types often have quite different spin–lattice relaxation times. For example, alkyl radicals are typically much more easily saturated than peroxy radicals. Thus, if both alkyl and peroxy radical spectra are superimposed, one can detect the presence of at least two different types of radicals by carrying out selective power-saturation measurements. Similar distinctions can be made between radical cations and radical anions, where one may have a much shorter spin–lattice relaxation time than the other. An example involves the photoionization of chlorophyll in vesicle systems with an electron scavenger such as a halogenated quinone present. Typically one detects the chlorophyll cation radical signal superimposed on the electron acceptor anion radical signal, but they can be distinguished by their different responses to microwave power saturation. The radical anion involving a halogen atom will typically saturate with much more difficulty than will the chlorophyll cation radical, which is more characteristic of an organic free radical.

Magnetic relaxation measurements can also be used to determine the spatial distribution of paramagnetic species in a solid matrix. This is important if the spatial distribution is nonuniform. Often when radicals or paramagnetic species are produced by photolysis or by radiolysis they are trapped in a nonuniform manner, and this can be detected by a careful analysis of the magnetic relaxation characteristics of the radical.

Potential and demonstrated applications of electron spin resonance are ubiquitous. The technique is particularly useful because it is sensitive only to those species that are paramagnetic. If these are important reaction intermediates, one has a selective analytical technique to look only at those specific types of reaction intermediates. Paramagnetic species are probably much more widespread than is generally believed. Radicals are typically reactive species, so appropriate time-domain or trapping techniques must be used to detect them before they decay into nonparamagnetic species. In such studies, electron spin resonance plays a major role.

XI. NEW DEVELOPMENTS

In the past few years electron spin resonance has seen impressive developments in the area of time-domain techniques. Improved solid-state microwave components, fast digitizers, and computer systems have made possible new types of pulsed electron spin resonance experiments. The sensitivity of these new experiments also makes them practical for many new chemical, physical, and biological problems. A second-generation general-purpose pulsed electron spin resonance spectrometer has been commercially introduced which makes these pulsed techniques, including pulsed ENDOR, widely available. Fourier transform and two-dimensional electron spin-echo techniques are also included. A newer commercial development is pulsed ESR and pulsed ENDOR at the higher frequency of 95 GHz. This has particular advantages of better sensitivity for small samples and enhancement of ENDOR sensitivity for low gyromagnetic ratio nuclei.

In continuous-wave electron spin resonance, extended multifrequency capabilities from 0.3 to over 100 GHz have been developed based on loop-gap and other types of resonators. The lower frequencies seem particularly useful for some biological applications. Very high frequency spectrometers have also been developed up to 700 GHz, with commercial instrumentation available at 95 GHz. The higher frequencies are based on Fabry-Perot resonators and give superior g-anisotropy resolution, suppression of second-order effects, and better sensitivity for small samples.

SEE ALSO THE FOLLOWING ARTICLES

Atomic Physics • Atomic Spectrometry • Chemical Kinetics, Experimentation • Nuclear Magnetic Resonance • Perturbation Theory • Quantum Mechanics

BIBLIOGRAPHY

Atherton, N. M. (1993). "Principles of Electron Spin Resonance," Ellis Horwood, London.

Berliner, L. J., and Rueben, J., eds. (1989). "Biological Magnetic Resonance. Spin Labeling: Theory and Applications," Vol. 8, Plenum, New York.

Dikanov, S., and Tsvetkov, Y. D. (1992). "Electron Spin Echo Envelope Modulation Spectroscopy," CRC Press, Boca Raton, Fla.

Hoff, A. J., ed. (1989). "Advanced EPR: Applications in Biology and Biochemistry," Elsevier, Amsterdam.

Keijzers, C. P., Reijerse, E. J., and Schmidt, J. (1989). "Pulsed EPR," North-Holland, Amsterdam.

Kevan, L., and Bowman, M. K., eds. (1990). "Modern Pulsed and Continuous-Wave Electron Spin Resonance," Wiley, New York.

Kevan, L., and Kispert, L. (1976). "Electron Spin Double Resonance Spectroscopy," Wiley (Interscience), New York.

Kevan, L., and Schwartz, R. N., eds. (1979). "Time Domain Electron Spin Resonance," Wiley (Interscience), New York.

Kurreck, H., Kirste, B., and Lubitz, W. (1988). "Electron Nuclear Double Resonance Spectroscopy of Radicals in Solution," VCH, New York.

Pilbrow, J. R. (1990). "Transition Ion Electron Paramagnetic Magnetic Resonance," Oxford Univ. Press, London.

Slichter, C. P. (1989). "Principles of Magnetic Resonance," 3rd ed., Springer-Verlag, New York.

Weil, J. A., Bolton, J. R., and Wertz, J. E. (1994). "Electron Paramagnetic Resonance," Wiley, New York.

Electrophoresis

S. P. Spragg
Birmingham University, England

I. Underlying Theory
II. Free-Solution Electrophoresis
III. Electrophoresis in a Stationary Matrix
IV. Electrophoresis in Oscillating Electric Fields

GLOSSARY

Chemical potential Ability to do chemical work; analogous to mechanical potentials.

Conductance Inverse of electrical resistance of an ionic solution. A specific conductance is defined as the inverse resistance of 1 M^3 of solution under a potential of 1 V.

Coulomb Unit of electrical charge having dimensions of ampere seconds. Bringing 1 mol of charges of magnitude ze (e = charge of an election, z = number of charges per particle) up to a potential V requires doing work on the charge equal to $zeVN$ (N = Avogadro's number); here, eN is called the faraday (F), having units of coulombs per mole.

Electrochemical equivalence Number of moles of an ion required to carry 1 F of charge.

Electrodes Conducting units inserted in a solution; negative electrode is called the cathode, and positive electrode is called the anode. A reversible electrode is produced by coating the metal with one of its salts, for example, silver/silver chloride.

Frictional coefficients Generally, these are introduced in hydrodynamics in order to ascribe a Newtonian friction to particles or molecules. A molar frictional coefficient F is the product of the particulate coefficient and Avogadro's number (see Mole).

Isoelectric point When organic ampholytes (also see Polyampholyte) have two covalently linked basic and acidic ionic groups, the isoelectric point equals the hydrogen ion concentration where the net charge is zero.

Isotachophoresis A procedure similar in many respects to isoelectric focusing, since it separates macroions through differences in charge rather than size (the important property for polyacrylamide gel electrophoresis).

Mobility Normalized velocity of a particle. In electrophoresis the mobility is the velocity per volt $\times$ meter ($m^2\ sec^{-1}\ V^{-1}$); this is different from the mobility coefficients found in diffusion and sedimentation.

Mole Chemical quantity used to describe a unit containing $\sim 6.03 \times 10^{23}$ molecules or ions; this constant is called either Avogadro's or Lochschmidt's number.

Polyampholyte Polymerized ampholyte in which some of the ionizable groups of the monomeric ampholyte remain free after polymerization to form a macroion. A protein is a common example and is formed by condensation of amino acids (called zwitterions) through the terminal carboxyl group of one with the terminal amino group of the next to form a peptide.

Solvent, solute A solvent is chemically identified as the substance present in considerable excess in a mixture, while the solute is that present in minor quantities.

ELECTROPHORESIS is the movement of charged particles when suspended or dissolved in a polar liquid and the mixture is placed in an electric field. This generally excludes the movement of charged particles in gases, although with modifications the basic concepts could be applied to these suspensions. Although the procedure is more widely known for studying biological molecules and colloidal particles, the underlying principles were discovered during early research on the electrical conduction of solutions of inorganic salts. For most purposes the movement of the ion is translational and carried out under the influence of a constant electric field. Movement in an oscillating electric field could be included within the general term "electrophoresis"; at high frequencies, however, the molecular motions of dipoles must be included, and here quantum treatments are more relevant, so these are not considered to be electrophoretic phenomena. For this reason little is mentioned in this article on the effects of oscillating fields on charged ions. It is also implicit in the considerations of electrophoresis that the particles or macroions are insulators and that the internal arrangements of ions in molecules are not affected by external fields.

I. UNDERLYING THEORY

A. Background

When a constant electric field is applied across a column of water, current flows and gases evolve at the electrodes—hydrogen at the cathode and oxygen at the anode. This phenomenon is known as electrolysis and was discovered by M. Faraday in 1832. He provided quantitative laws to relate the amount of chemical decomposition with the number of coulombs passed. It is in recognition of his work that the primary electrochemical constant is called the faraday. It was not until much later that the source of the decomposition was discovered, since the understanding of this phenomenon required the knowledge that salts can dissociate into ions when dissolved in a polar solvent, a concept first appreciated by S. Arrhenius in 1887. The dissociation of salts into ions is now universally accepted, and the fact that dissolving a salt in water causes the spontaneous dissociation of a molecule into one or more cations (positively charged) and anions (negatively charged) is never questioned. The driving force for this dissociation arises from the differences in energy between solvating the free ion and the energies of interaction of the intact molecule with the solvent. It is clear in chemical thermodynamics that these differences can be discussed through chemical potentials of the two states, but *ab initio* calculations of these processes are not possible.

The experimental basis for these laws of conduction of solutions involved the measurement of conductivity of a solution by applying Ohm's law to the electrical measurements. From these experiments Kohlrausch showed that at infinite dilution each ion contributed a definite amount to the conductivity irrespective of the nature of the other ion. In order to explain these phenomena it became necessary to introduce the concept of electrochemical equivalences showing that the conductance of a solution is the product of the number of ion in the solution, the charge carried by each ion, and the velocity or their mobilities $\mathbf{u}_{\pm}$.

Quantitative relationships were developed between the current $\mathbf{i}$ carried through the solution by C_{+} gram moles of a univalent anion and the mobilities observed in an electric field $\mathbf{E}$:

$$\mathbf{i}_{+} = \mathbf{u}_{+}C_{+}\mathbf{E}. \tag{1}$$

The equivalent expression for the cation is the same as Eq. (1), and so the total current $\mathbf{i}$ is

$$\mathbf{i} = \mathbf{i}_{+} + \mathbf{i}_{-} = (\mathbf{u}_{+}C_{+} + \mathbf{u}_{-}C_{-})\mathbf{E}, \tag{2}$$

which can be compared directly with Ohm's law as used to define resistance,

$$\mathbf{E} = 1/k, \tag{3}$$

where $\mathbf{k}$ is the conductance, hence, $\mathbf{k} = \mathbf{u}_{+}C_{+} + \mathbf{u}_{-}C_{-}$ and the conductance is proportional to the velocities of the ions. The quantity of positive electricity carried in one direction by the cations is proportional to $\mathbf{u}_{+}$ and C_{+}, while that in the other direction by the anions is proportional to $\mathbf{u}_{-}$ and C_{-}. This makes it possible to define a coefficient that proportionates the current passed between the individual ions. This is called a transference or transport number:

$$\mathbf{T}_{+} = \frac{C_{+}\mathbf{u}_{+}}{C_{+}\mathbf{u}_{+} + C_{-}\mathbf{u}_{-}}$$

and

$$\mathbf{T}_{+} = \frac{C_{-}\mathbf{u}_{-}}{C_{+}\mathbf{u}_{+} + C_{-}\mathbf{u}_{-}}. \tag{4}$$

Here, $\mathbf{T}_{\pm}$ is dimensionless, but $C_{\pm}\mathbf{u}_{\pm}$ has dimensions $\mathrm{mol \cdot A \cdot sec^{2} \cdot liters^{-3}}$, making this equivalent to the number of molar coulombs required to sweep out a unit volume per second or number of faradays required to sweep out a unit volume per second.

In general, the parameter measured in electrophoresis is the velocity of the ion, which is done by identifying

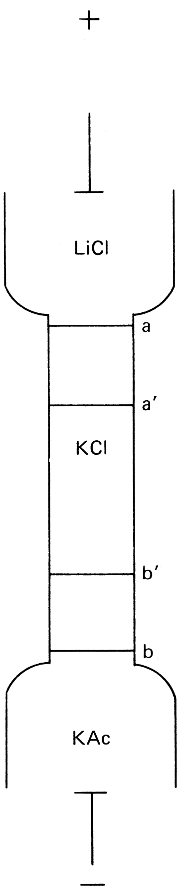


FIGURE 1 Idealized electrophoresis experiment in which boundaries between three salts are formed initially at *a* and *b*. After passage of a known quantity of electricity, the boundaries have moved to *a*′ and *b*′.

it by forming a suitable boundary between the ion and the solvent and measuring the velocity of this boundary. An elementary arrangement for making this experiment is given in Fig. 1. A solution of a simple salt, KCl, separates two solutions of other salts, LiCl in the anode compartment and potassium acetate in the cathode compartment. This produces initially two boundaries at *a* and *b* (note that the concentrations of the salts must be arranged so that the least dense solution is at the top). In choosing these salts it is important that the velocity of Li^+ be less than that of K^+ and the velocity of Ac^- be less than that of Cl^- if two sharp boundaries are to be maintained throughout the experiment. When a voltage is applied across the electrodes, boundary *a* moves to a' and *b* moves to b', so the distances aa' and bb' represent the velocities. Since the solution of KCl is homogeneous at the beginning, its conductivity is constant and the fall in potential between *a* and *b* will be uniform, so the two ions move through the same voltage gradient. Thus, $aa'/bb' = \mathbf{u}_+/\mathbf{u}_-$:

$$\therefore \mathbf{T}_+ = \frac{\theta_+}{\mathbf{u}_+ + \mathbf{u}_-} = \frac{aa'}{aa' + bb'}$$

and

$$\mathbf{T}_- = \frac{bb'}{aa' + bb'}. \tag{5}$$

If this experiment were performed, the sharpness of the boundaries would be affected by the concentrations of the three salts [see Eq. (4)]. A sharp boundary would form only if Li^+ did not overtake K^+ and Ac^- did not overtake Cl^-; at the limit this means that the velocity of Li^+ would be the same as that of K^+ and the velocity of Ac^- would equal that of Cl^-. For a set of these boundaries, Kohlrausch defined a regulating function for defining sharpening conditions, namely, when $\mathbf{T}_+/C_+$ is equal on both sides of the boundary for each ion of the same sign, a sharp boundary forms. This function can be derived from Eq. (4) if one remembers that the sum in the denominator represents the total conductance; hence, the mobilities must be equal if $T_+^{(1)}/C_+^{(1)} = T_+^{(2)}/C_+^{(2)}$; this condition is enlisted to give stacking of proteins in gel electrophoresis (Section III.D).

Boundary experiments were employed during the early part of the 20th century for studying aqueous salt solutions, and the rules that emerged are associated with such people as Hittorf and McBain. The results showed that ions were hydrated not equally, but by an amount that appeared to be linearly related to the size of the ions. Most of these studies were made using inorganic ions, and it was not until the 1920s that electrophoresis became a recognized tool for studying macroions such as those found in biological and colloidal systems. The impetus for this development came from the Uppsala school, where The Svedberg was developing an analytical ultracentrifuge while A. Tiselius was developing the successful electrophoresis procedure that bears his name. With centrifuges it was possible to recognize individual components in a mixture on the basis of mass, while with electrophoresis similar separations were made through differing charges and size. The inclusion of macroions in electrophoretic studies formed the foundation of modern electrophoretic techniques. These are used almost exclusively for studying biological macroions.

B. Moving-Boundary Electrophoresis

Without questioning the origin of charges on particles but assuming that macromolecules and particles carry either positive or negative charges, we can state that once an external electrical field is applied the ions drift to the electrode of opposite charge. There occurs a finite but vanishingly small time after the voltage is connected and before the current has reached its constant level. This time is determined by the electrical time constant of the

circuit plus the time required for the randomly moving molecules to adopt a directional motion (this takes altogether $<10^{-6}$ sec). This delay is not considered here because the electrophoretic mobility is slow relative to this transition. For most purposes it can be assumed that the current is constant throughout the experiment and the macroions move at a uniform velocity (dx/dt) determined by the field strength and total charge Q on the ion. In order to maintain a constant velocity, Newton's laws show that the movement in one direction must be opposed by an equal but opposite force. This opposing force is ascribed to frictional forces f between the macroions and the stationary solvent. (In fact, the solvent is stationary only at distances far removed from the surfaces of the particle.) The value of the frictional force is determined by the intrinsic size and shape of the ion, as well as the viscosity of the solvent, and increases with the velocity of the ion. Thus, we can equate the two opposing forces to give

$$Q\mathbf{E} = \frac{dx}{dt} f \quad \text{or} \quad \frac{dx}{dt} = \frac{Q\mathbf{E}}{f}. \tag{6}$$

If $\mathbf{E}$ is measured as volts per meter, then dx/dt is called the mobility of the ion for the chosen experimental conditions. The coefficient f has a theoretical foundation in hydrodynamics, and a functional relationship between f and the coordinates of the particle can be derived for a few regular shapes (for a sphere it is known as the Stokes' equation, but there are mathematical solutions for ellipsoids and cylinders). It is also known that f is inversely proportional to the randomizing effects of diffusion of a large number of ions (Einstein–Sutherland relationship). Essentially, Eq. (6) describes the movement of a single ion under the influence of an electrical field. Rarely, if ever, can one ion be studied experimentally, because at finite concentrations of ions there are $>10^{13}$ ions per liter (a 10^{-10} M solution of 0.1 μg liter^{-1} for an ion of relative mass 10^3 contains 10^{13} ions per liter). Diffusion of this population of ions spreads the boundary about an electrophoretically transported point called the centroid, and it is the velocity of this point that is described by Eq. (6) for experimental situations. The centroid or first moment $(\bar{x})$ can be evaluated from Eq. (7) using a set of rectangular coordinates determined experimentally over an electrophoresing boundary,

$$\bar{x} = \frac{\int_{x_1}^{x_2} xy\ dx}{\int_{x_1}^{x_2} y\ 2dx}. \tag{7}$$

The shape of the boundary approximates a Gaussian profile, but the exact description of the shape depends on the starting conditions of the experiment, the mobilities of the various ions, and whether sharpening of the boundary occurs as a result of electrical effects (Kohlrausch regulating function; see Section I.A). If the conditions are such that these factors do not distort the boundary and the starting condition was similar to that described for Fig. 1, where the concentration of the ion is uniform on one side of the boundary and zero on the other, then the boundary shape can be described by Eq. (8):

$$C(x) = \frac{1}{2} C_0 \left[1 - \frac{2}{\sqrt{\pi}} \int_0^{(\pi/2)(Dt)^{1/2}} \times \exp\left(- \frac{x^2}{2\sqrt{Dt}} \right) dx \right]. \tag{8}$$

In Eq. (8), $C(x)$ is the concentration at x after t seconds of a substance having a diffusion coefficient D (the starting concentration was C_0). If, however, a zone of width h containing the macroion is introduced between the electrode solution and the main bulk of the solvent, both sides of the zone diffuse to produce a double sigmoid shape, which at its simplest can be described by Eq. (9):

$$C(x) = \frac{C_0}{2} \left[\mathrm{erf}\left(\frac{h - x}{2\sqrt{Dt}} \right) + \mathrm{erf}\left(\frac{h + x}{2\sqrt{Dt}} \right) \right], \tag{9}$$

where $[\mathrm{erf}\,(z) = (2/\sqrt{\pi}) \int_0^a \exp\,(-a)\, da]$ and the distribution is symmetrical about a plane at $x = 0$. With both types of boundary, the centroid [$\bar{x}$, Eq. (7)] corresponds to the center of the boundary, but if the boundary is warped by electrical inhomogeneities or contains a mixture of unresolved ions of slightly differing mobilities, then $\bar{\mathbf{x}}$ will still correspond to the required centroid of the boundary but not necessarily its geometrical center. This explains why it is important to use the centroid for calculating average mobilities of electrophoresing boundaries.

The diffusion coefficient D and frictional coefficient f [Eqs. (6), (8), and (9)] of an ion are similar to that found for a neutral molecule. However, because the salt dissociates when dissolved, but electroneutrality must be maintained throughout the solution despite each ion having different diffusion rates, it is necessary to modify Eq. (6) (and the other equations where a diffusion coefficient is employed to replace frictional forces). This is done by replacing the frictional coefficient in Eq. (6) by $\mathrm{F}(\mathbf{u}_+ + \mathbf{u}_-)/2\mathbf{u}_+\mathbf{u}_-$ ($\mathbf{u}_+$ and $\mathbf{u}_-$ are the mobilities of the anion and cation, respectively) to give the Nernst equation.

In any solution of ions there always occurs electroneutrality, so every positive ion has a counter negative ion. For macroions these are called gegenions. They can be small ions such as Na^+ or Cl^- or larger organic ions. Each ion transports a proportion of the current, and since small ions have greater mobilities than macroions, their transport numbers dominate the system, which means that they carry most of the current on both sides of an interface. The necessity of maintaining electroneutrality throughout

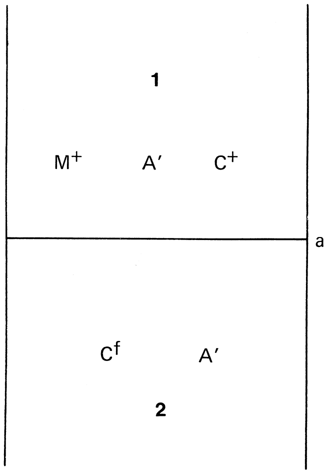


FIGURE 2 Equilibration of ions across a boundary (*a*) containing the macrocation M and two small gegenions A′ and C+.

the solution means that concentrations of individual ions are not equal across a boundary containing a macroion. This inequality is called the Donnan phenomenon (after an Irish electrophysiologist who first discussed it). It can be explained using the system given in Fig. 2, where a macrocation is present on only one side of the boundary. Thus, at equilibrium there occur the following equalities (brackets indicate chemical activities, which in dilute solutions equal concentrations):

$$\begin{aligned} \text{Phase 1:} \quad & [M^+] + [C^+] = [A^-], \\ \text{Phase 2:} \quad & [C^+] = [A^-]. \end{aligned} \tag{10}$$

At equilibrium the chemical potential of the solvent in phase 1 must equal that in phase 2 while $[C^+]_1 < [C^+]_2$ and $[A^+]_1 > [A^-]_2$, so the effect is an intrinsic instability at a free boundary. If this boundary were formed across a physical membrane permeable to A^- and C^+ but not M^+, an electrical potential and osmotic pressure would exist across the membrane. This hydrostatic pressure cannot occur in free-solution boundaries. Clearly, the extent of the instability is considerably reduced if the molar concentrations of the gegenions are high relative to that of the macroion. This condition is normally employed in electrophoresis of macroions. It should be mentioned that the generation of an osmotic pressure across a membrane separating phase 1 and phase 2 is normal for all molecules whether they are neutral or charged. The effect of these instabilities is to produce a rapid radjustment of concentrations of ions immediately after a boundary is formed, making the sum of electrochemical and chemical potentials equal across the boundary. It is to reduce these disturbances in free-solution electrophoresis with macroions that the solvent is dialyzed to equilibrium against the solution before an experiment is begun.

With respect to the physical factors influencing the stability and shapes of boundaries in electrophoresis, it is apparent that all ions in the solution are moving toward their oppositely charged electrodes. Naturally, various physical properties are employed to detect and identify the ions (e.g., pH indicators could be used to study the movement of H^+ and absorption of ultraviolet light for proteins or nucleic acids). The question arises, however, as to how many boundaries would form when a solution of ions were electrophoresed. As a result of the work of Longsworth and Dole it is possible to define for any system the number of boundaries. Their work shows that for most purposes in a system containing n ions there will occur a maximum of $n - 1$ boundaries. If there are p anions and q cations, then $q - 1$ boundaries move toward the cathode, $p - 1$ boundaries move toward the anode, and one stationary boundary forms. A stationary boundary does not necessarily mean that no ions are being transported. It means that there are no visible indications of this transport. Hence, a stationary boundary will form if the transport number of the ion is the same on both sides of an interface, which means the ion moves at the same velocity on both sides or the concentrations are equal across the interface [Eq. (4)]. In the case of proteins the greater proportion of the charge of the ion arises from dissociation of the exposed side chains of the condensed amino acids, so the total charge is determined by the relative amount of dissociation of these groups, which in turn is controlled by the pH of the solution. Thus, for all zwitterions there is a pH where the total charge is zero (isoelectric point) and the molecule has zero mobility. At this point it is possible to produce a stationary boundary that does not transport ions, but this is different from a stationary boundary induced by the electrochemical effects described by the Kohlrausch regulating function. Producing the stationary boundary by isoelectric focusing is discussed in Section III.B.

C. Origin of Molecular and Particulate Charge

Equation (6) is a general expression for movement of a single charged particle in an electric field, and it is possible to relate the charge to other molecular parameters. It can be shown from irreversible thermodynamics that the flow of a mole of ions (J) in a system is described by Eq. (11):

$$J = L_0\mathbf{E} + \sum_i L_i\,(\partial \mathbf{u}_i/\partial x), \tag{11}$$

where $\mathbf{E}$ is the electric field strength, $(\partial \mathbf{u}_i/\partial x)$ is the force caused by concentration gradients of any ionic or neutral species in the solution [for many purposes this mass transport is expressed by the diffusion, Eq. (8)], and L_0 and L_i are phenomenological coefficients. The latter are defined by molecular parameters and composition of the

solution but do not depend on the magnitude of the forces, while the flow J is directly proportional to the mobility $\mathbf{u}$ [Eq. (6)]. No existing theory of electrophoretic mobility is consistent with Eq. (11), since most assume that $\mathbf{u}$ or J is proportional only to the electric field strength. In any analysis it is assumed that the fluid is stationary at the boundary of the particle, and fluid movement is equal but opposite in direction to particulate movement when viewed at distances far removed from the particle. The result of this definition is to consider the mobility as a vector. Using these mathematical boundary conditions, it was possible to solve equations formulated using Eq. (11) and to include the close interaction of gegenions with the charged spherical particle (Debye–Hückel relationship) and viscosity η, so giving the approximate relationship

$$\mathbf{u} = \frac{Q}{6\pi\eta R}\left(\frac{1+\kappa r_i}{1+\kappa a}\right)P(\kappa a) \qquad (12a)$$

$$\kappa = \left(\frac{8\pi N\varepsilon^2}{1000\bar{D}kT}\right)^{1/2} I^{1/2}. \qquad (12b)$$

In these equations r_i is the radius of a typical gegenion, R is the radius of the macroion, $a = R + r_i$, $\bar{D}$ is the dielectric constant of the medium, k is the Boltzman constant, N is Avogadro's number, ε is the charge on an electron, and $I = 0.5\sum_i Z_i^2 C_i$ (the ionic strength of the solution, where Z is the valency of the ith ion of concentration C_i); $P(\kappa a)$ is a dimensionless function required to allow for the effect I has on the effective radius of the particle and has values that vary between 1.0 and 1.5. The inclusion of a in Eq. (12) is necessary in order to include the effect that the close approximation of the neutralizing gegenions has on the particle. The effect is to increase the radius of the macroion beyond that expected from the neutral molecule. Several workers have changed Eq. (12) to allow for other electrical phenomena, but despite these efforts it is reasonable to assume that no relationship between electrophoretic mobility and molecular parameters is rigorously applicable to macroions.

This conclusion is disappointing because the major reason for the development of electrophoresis was to relate the charge of colloidal suspensions and biological macromolecules to their known molecular parameters. The use of electrophoresis in recent times has been restricted to empirical studies where electrical forces have been used to separate mixtures of charged particles or macroions into individual components.

The technique has many applications and has even been used for separating suspensions containing apparently neutral molecules, because all particles in stable suspensions carry a net charge. The charging of a colloidal particle is necessary if precipitation is to be reduced. This produces a balance between the repulsive forces produced by like charges and the ubiquitous attraction between all molecules caused by dispersion forces. In effect, the particles are held apart in a low potential energy well similar in shape to that found for all intermolecular forces (called the Lennard–Jones potential). Thus, it is reasonable to ask how a particle formed from an apparently neutral molecule can become charged in suspension. (An example of such a suspension is a suspension of carbon particles to give printer's ink.)

The source of the charges is adsorbed ions from the solvent collecting on the exposed surface of the particle. The energy for this adsorption comes from the interfacial energy found when two phases share a common boundary and is called surface excess energy (or for aqueous solutions with respect to air, surface tension). A thermodynamic argument can be used to describe the energy of adsorption of a solute. This includes the difference between the chemical potentials of the dissolved ions in the occluded solvent around the particle and those in the bulk phases. This provides the basis for the Gibbs adsorption isotherm, which shows that the surface excess concentration of the solute is proportional to the logarithm of the chemical activity of the solute in the bulk solutions (this is equal to concentration for dilute solutions). The origin of this energy is the differing interactions of the ions with the solvent in the bulk of the two phases and that far removed from the interface. These interactions lead to either an accumulation or a deficit at the interface.

A natural outcome of these phenomena is that the charge measured by electrophoresis of macroions or particles is not necessarily equal to that found by algebraically summing the ionizations of the intrinsic side groupings of the macroion. (This can be estimated for acidic and basic groups by chemical titration.) The term "ζ potential" has been used to describe this total electrophoretic charge. It is the charge at the plane of slip between the charged macroion or particle and the bulk solvent. The spatial dependence of the charges depends on many factors, but an idealized example of the distribution is shown in Fig. 3 for a positive ζ potential. It should be mentioned that the distribution shown in Fig. 3 is an average because there is a general randomizing movement of the ions (Brownian motion). This means that the instantaneous relationship for individual ions need not equal that shown in Fig. 3, and it is still possible for collisions to occur between the particles despite the fact that on average they all carry the same charge. These collisions eventually precipitate the suspensions.

A ζ potential is more generally used in discussions of colloidal suspensions (e.g., pigments in paints) and is not often applied to the macroions found in biology. Here, the major part of the charge arises from ionization of the polar groups of the condensed subunits in the polymer. For proteins (an example of polyampholytes) there are ionizable carboxylic, amino, phenolic, and mercapto groups,

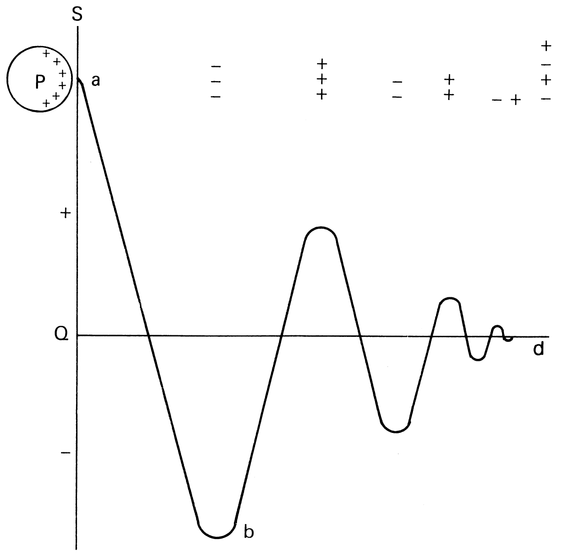


FIGURE 3 Debye–Hückel distribution of spatial charges Q along d from a solvated particle P having a layer of solvent S which contains positively adsorbed ions. The total potential between points a and b is called the ζ potential.

and all show a dependence of the proportion of ionization on the pH of the medium. Free carboxylic groups are fully ionized at pH > 5, while amino groups form quaternary ammonium ions at pH < 10. The other ionizable groups in the protein fall within this range. The isoelectric point for all proteins is not constant, but in general it falls between pH 4.5 and 8. For nucleic acids the dominant ionizable group is the free phosphate, which ionizes at pH < 2. (There are two free ionizable groups on phosphate, and the second ionizes at a much higher pH.) The amino groups of pyrimidines and purines contribute insufficient charge to neutralize the phosphate, so at pH > 3 the nucleic acids are negatively charged. Some polymeric carbohydrates (e.g., starch and cellulose) are mainly neutral, although residual covalently linked phosphates remain from the enzymatic condensation of glucose 1-phosphate to form starch. Others (e.g., agar) are strongly charged due to preserved covalently linked sulfate and other ionizable groups. There also occur oligomeric carbohydrates containing more than one type of sugar, and these carry both amino and sulfated sugars in the chain, which endow the oligomers with charge. These charged oligomers are often condensed to proteins and lipids, moving the net charge to regions outside the expected isoelectric range for pure proteins and lipids. All these groups can be titrated by acid using pH to monitor the progress of the reactions, and from the results the net charge at each pH can be calculated. These charges are seldom equal to those obtained by measuring mobilities free in solution at different pH values except at the isoelectric point. (Often this is taken as zero by definition in order to calculate the other charges.) For example, the egg protein ovalbumin has an isoelectric point of pH 4.5, but at pH 7 the number of charges determined from pH titration was about −17, while that from the mobilities was −10; the difference is thought to be caused by preferential adsorption of Cl^-.

One major contribution made by electrophoretic experiments with proteins was the measurement of interactions between peptide chains of complex proteins (such as hemoglobin) with other ions, as well as their own self-association. Most biologically active proteins are formed from aggregates of peptide chains (the subunits), and these chains combine through a variety of weak intermolecular forces such as salt links or van der Waals dispersion forces. If the energies of these intermolecular forces fall within a range similar to that arising from thermal activities, the aggregates dissociate on dilution to produce an equilibrium mixture containing free subunits plus the aggregate. Electrophoresis provided one of the two experimental procedures for studying the energies of interaction between the subunits.

The importance to biology of these studies can be illustrated through the example of hemoglobin, which transports four molecules of oxygen, each having different energies of interaction between oxygen and a heme. Hemoglobin is made from two types of subunits (two of each subunit per molecule of hemoglobin), and the four energies of interaction with oxygen arise from conformational changes in the subunits, which occur as each molecule of oxygen is added. The results, which were used to evaluate the sequence of changes in solution, were obtained from ultracentrifugal and electrophoretic experiments. The experimental expression of the dissociation was a distortion of the moving boundary caused by increasing amounts of the dissociated subunits at the lower concentrations of the macroions in the diffusing boundary. Furthermore, the mobility of the distorted boundary [estimated from the centroid, Eq. (7)] was dependent on the starting concentration of protein, because the various proportions of subunits affected the averaged mobilities. Mathematical models describing these results contain equilibrium constants that describe the dissociation. These can be numerically evaluated by fitting the models to the boundary shapes and dependence of mobilities on concentration. The models of Cann and Gilbert have been used to interpret a variety of experimental effects caused by the association of ions.

D. Stationary Boundaries

Apart from the origin of charges, the earlier part of this discussion was concerned with nonequilibrium transport of ions. This is obviously an important situation when one is describing the conductivity of solutions and separating components from a mixture, but it is possible to arrange

experimental conditions so that no net transport of ionic species occurs. It is obvious that transport of a solute can occur only if the density of the ion is greater than that of the solvent. Despite the considerable difference between electrical and gravitational forces, it would not be possible to generate a sufficiently large electrical field in polar solutions that would force a particle to settle in a solvent of higher density than itself. The current would be sufficient to boil the solution. Hence, it is possible to design an experiment where at the beginning a band of solution containing macroions is placed on top of a solvent column consisting of a preformed density gradient. This gradient can be formed by varying the concentration of a neutral molecule down the column. (Sucrose might be used since it is neutral and has a density of $\sim$1.6 g cm^{-3}; protein densities are $\sim$1.3 g cm^{-3}.) If the band is now electrophoresed down the column, a point occurs where the density of the macroion is less than that of the solvent; transport stops and a stable boundary forms in this plane, which is often called its isopycnic point. The charge on the macroion has not been neutralized by the gradient, so this is not an electrical equilibrium but an equilibrium between two equal but opposite forces on the macroion.

A different situation can be generated for polyamopholytes where, instead of electrophoresing a band along a density gradient, one forces the band to travel along a pH gradient. In this case the net charge of the polyampholyte decreases as it moves toward its isoelectric point until it reaches the pH where it carries no net charge and a stationary boundary forms as a result of the equilibrium between electrical and diffusive forces. An infinitely thin zone is not formed at equilibrium for either of these conditions because diffusion disperses the zone and the zone is stable only as long as the electrical field is applied. The resulting shape is approximately Gaussian. A functional relationship can be derived for the isolectric equilibrium that relates the concentration at any point within the zone to the molecular properties of the macrion.

The electrophoretic force qV on the polyampholyte in an electric field of V volts per unit distance is negative because it moves against the voltage gradient and is

$$-qV = (dx/dt)Nf. \tag{13}$$

The opposing force at equilibrium is given by Fick's first law of diffusion for the mass flux dm/dt through an area A:

$$dm(D)/dt = -DA\,dc/dx. \tag{14}$$

At equilibrium the net transport across any plane is zero, so the electrophoretic flux equals that of diffusion. The electrophoretic mass flux $dm(e)/dt$ is

$$dm(e)/dt = -ACqV/Nf. \tag{15}$$

Equating Eq. (14) with Eq. (15) and eliminating A gives

$$d\ln C/dx = -qV/DNf. \tag{16}$$

The net charge q of the polyampholyte is a function of pH, and if we assume a linear pH gradient, then the charge at position $x[q(x)]$ is defined as

$$q(x) = q(0) + (x - x_0)\,dq/dx, \tag{17}$$

where dq/dx is determined by dq/dpH and dpH$/dx$, the experimental variate. The reference charge $q(0)$ in Eq. (17) is chosen as that found at the isoelectric point of the macroion (x_0), which is zero. Thus, substituting Eq. (17) into Eq. (16) gives ($F \equiv Nf$)

$$\frac{d\ln c}{dx} = -(x - x_0)\frac{dq}{dx}\frac{V}{FD}. \tag{18}$$

When values of $x - x_0$ are small, dx can be replaced by $\frac{1}{2}d(x - x_0)^2$. Thus, retaining the experimental variable dpH$/dx$, Eq. (18) becomes

$$d\ln c = -\frac{dq}{2d\text{pH}}\frac{d\text{pH}}{dx}\frac{V}{FD}\,d(x - x_0)^2. \tag{19}$$

Integration between the limits of c and x by defining c_0 as the concentration at x_0 gives

$$C = C_0 \exp\left[\frac{-(dq/d\text{pH}\,(d\text{pH}/dx)\,V(x - x_0)^2}{2FD}\right]. \tag{20}$$

By analogy with the Gaussian probability relationship, the width of the profile equals $[FD/(dq/d\text{pH})(d\text{pH}/dx)V]$, where $(d\text{pH}/dx)$ and V are two experimental variables, so the only molecular parameters are $(dq/d\text{pH})$, F, and D. The product FD equals RT (R being the gas constant and T being the temperature). Hence, Eq. (20) becomes

$$C = C_0 \exp\left[\frac{-(dq/d\text{pH}\,(d\text{pH}/dx)\,V(x - x_0)^2}{2RT}\right], \tag{21}$$

in which the numerator is an energy term. A similar expression can be derived for the isopycnic experiment. However, the width of the band would be inversely proportional to the density gradient and density of the macroion instead of being inversely proportional to the pH gradient and differential charge for equilibrium at the isoelectric point; the latter is called isoelectric focusing.

E. Electrophoretic Mobility, Sedimentation Coefficient, and Diffusion Coefficient for Macroions

Three common mobilities are used to describe the size and shape of a macroion: (1) electrophoretic mobility,

which describes charge and shape; (2) sedimentation coefficient s, which describes mass, density, and shape; and (3) diffusion coefficient D, which describes shape only. The common property for these three is shape, which in turn is contained within the formulation of the frictional coefficient f [as in Eq. (6)]. Thus, each mobility can be related to the other two, provided that the frictional coefficient is identical for the three conditions. Experience has shown that this is a valid assumption, provided that experimental conditions are constant during the determinations (constant composition of solvent, pH, ionic strength, temperature, and concentration of macroion). Using the ideal relationships for s, D, and $\mathbf{u}_{\pm}$ [Eqs. (22)–(24)], the relationship for mobility and sedimentation is given in Eq. (25):

$$s = \frac{M(1 - \rho_{\mathrm{s}}/\rho_{\mathrm{m}})D}{RT} \tag{22}$$

$$D = RT/F \tag{23}$$

$$\mathbf{u}_{\pm} = Q_{\pm}/F \tag{24}$$

$$s = \frac{M(1 - \rho_{\mathrm{s}}/\rho_{\mathrm{m}})\mathbf{u}_{\pm}}{Q}, \tag{25}$$

where M is the molar mass, ρ_{s} is the solvent density, ρ_{m} is the macroion weight density, and Q is the charge of macromolecule calculated from the product of the number of charges and the charge on an electron. Within the maximum allowable experimental conditions that can be used in sedimentation, this procedure gives velocities that are about 10 times that produced by electrophoretic forces. This fact explains in part why for many pruposes sedimentation has proved a more usable preparative technique for resolving mixtures than electrophoresis in free solution.

II. FREE-SOLUTION ELECTROPHORESIS

A. Conventional Procedures

Until the late 1950s electrophoretic experiments were carried out in columns of aqueous solutions. The equipment was, in principle, that given in Fig. 1, but, in practice, it consisted of a glass ∪-tube having a square cross-section and constructed from three sections (divided across the channel of the ∪). Each section carried parts of both limbs of the ∪-tube, and they were designed so that each could slide across the others in a plane set normal to the direction of the channels in the ∪-tube. The top section had outlets from the ∪-tube in order to connect the limbs to separate electrode vessels, while the bottom section was essentially a connector to complete the bottom part of the ∪-tube. The middle section was the optical component, and it could be divided into two sections, but it is more convenient to describe the action as though the middle were a single section. The middle carried the optical windows on two sides of each channel, and it was usual to monitor the movement of the boundary through these windows by non-invasive optical methods. The experiment was performed by first introducing the solvent into the cell while all the sections were connected. Then one would slide the center section plus the top section across the bottom, so isolating the bottom section. The solvent was removed from one of the limbs in the ∪-tube and replaced by the solution (dialyzed if the experiment involved macroions). The top section was then moved back, leaving all the sections isolated from the one limb and replaced by solvent before the connection was made to the electrode vessels. The whole assembly was mounted on a mechanical support, and the solvent was added in order to fill the electrode vessels and top of the cell. Saturated KCl solution was added to the bottom of each electrode vessel after the electrodes were inserted. The complete unit was placed in a thermostatted bath having optical windows for examining the center section from outside the bath. When it was equilibrated for temperature, the middle section was moved across to complete the channel through the ∪-tube. This operation produced a boundary between solution and solvent at the two interfaces between top, middle, and bottom sections. During the experiment, the movement of the boundary was observed by a variety of methods of which the most popular wer schlieren and interference optics. The patterns were recorded on films, which were eventually measured to calculate velocities and boundary profiles. Considerable care was taken before the experiment began to equalize the column heights, thus reducing hydrostatic distrubance. The electrodes were made from silver wire and coated with AgCl, so when the electrodes were immersed in the saturated KCl solution at the bottom of the electrode vessels the major current was transported by Cl^-. Electrolysis did not occur, provided that the solution under investigation contained K^+ and Cl^-. The assembly was vertically mounted, and the densities of the solutions had to increase from the top to the bottom in order to minimize mechanical mixing of the boundaries.

Electrical heating of the solution occurred during the experiment, and the resulting density changes could cause convection, which would modify the boundaries and in extreme cases destroy them. To reduce this convection, experiments were normally made at the temperature of maximum density of water (∼277°C) while the current was kept low by employing organic solutes to buffer the pH when solutions of polyampholytes were examined. Despite all these precautions, considerable experience was required in setting up the experiments and interpreting the results. This equipment (colloquially called the Tiselius apparatus) was used extensively by biophysical chemists when

TABLE I Examples of Mobilities u of Several Proteins Measured in Solutions Having Ionic Strengths of ∼0.1 M

Protein	Isoelectric point	pH	$u \times 10^{-9}$ ($m^2\ sec^{-1}\ V^{-1}$)
Ovalbumin	4.58	6.8	−6.1
α-Lactoglobulin A	5.09	5.3–6.0	−0.63
α-Lactoglobulin B	5.23	5.3–6.0	−0.12
α-Casein	4.1	8.6	−6.7
Serum albumin	4.7	8.6	−6.7

studying mixtrues of proteins. Information from these experiments was obtained on the role of charges in stabilizing macroions in solutions. Some selected results of mobilities and isoelectric points are given in Table I in order to gauge the magnitudes of the mobilities as well as illustrate the values of typical isoelectric points (where zero mobility would be observed) for these proteins.

It should be pointed out that results from free-solution electrophoresis were tested against models used to generate Eqs. (6)–(12), and these comparisons showed that no general model could be proposed to describe adequately the properties of charged macroions in solution. When this conclusion is combined with the technical problems of interpreting incompletely separated boundaries and the difficulties of covering wide-ranging conditions for the experiments, it is clear why the use of the Tiselius apparatus declined during the 1960s and hydrodynamic properties were studied by ultracentrifugal analyses. The ultracentrifuge had a built-in stabilizing force that reduced convection, and it could be used to study neutral molecules as well as charged molecules in a wide variety of solvents and temperatures. Possibly because of its unique features, electrophoresis did not disappear but evolved into a qualitative tool through the use of stabilizing gels. This made the technique one of the most widely applied procedures in biochemical studies (see Section III). As for small ions, the understanding of the conductance of ionic solutions has not been seriously pursued in recent times, so the need to develop new experimental procedures that use boundaries has not been present.

B. Electrophoretic Light Scattering

A development in light scattering in the 1970s made it possible to study the electrophoretic movement of macroions without forming boundaries. The physical principle behind the technique was to measure molecular motions through the Doppler shift in the frequency of scattered light relative to the incident beam. The measurements were made using a single-mode laser as the source of light. When light is scattered from a stationary object, its frequency is the same as the incident light, but in solutions the molecules become stationary only when the temperature is absolute zero. At higher temperatures molecules show the random motion known as Brownian motion, and light scattered from these is shifted to higher or lower frequencies depending on the relative direction the molecules were moving at the time of scatter. This is called quasi-elastic light scattering (QEL) in order to contrast it with the situation where no change in frequency of the light occurs during scattering, which is called elastic scattering. The frequency shift is less than 1 MHz for macroions and becomes smaller as the size of the molecule increases. Experimentally, the problem was to determine this small shift relative to the incident light frequency of about 10^{15} Hz.

Measuring small shifts in frequency of light can best be achieved by interferometric methods in which the scattered light is mixed with the incident light at the surface of a detecter (e.g., a photomultiplier cathode). The resulting signal has a frequency that equals the difference between the two frequencies. Two basic procedures have been developed for recording these beat frequencies: heterodyne and homodyne detection. The heterodyne method requires either a direct mixing of the incident light with the scattered light (after reduction of the incident intensity) on the detecter surface or the positioning of a stationary scatterer in the solution (e.g., a captive polystyrene sphere of considerably larger dimensions than the macroions being investigated) and recording the resultant signal. In homodyne experiments the scattered light at time t is autocorrelated with that recorded a short time later $(t + \tau)$. Unlike conventional light scattering, where the light intensity is recorded and which requires high incident light levels, the intensity of scattered light for QEL must be sufficiently low that the photons arriving at the detecter can be counted. When counting is employed it is possible to sample the flux of photons temporally, and using statistical procedures of autocorrelation the counts are processed to produce a relaxation curve with an exponential decay of the first-order correlation function with time. The time constant of the curve is related to the diffusion constant of the scattering macroions and therefore provides a measure of the diffusion of the ions. These measurements are obtained from a solution that is at true equilibrium, and no separation occurs during the experiment.

Electrophoretic applications of this method depend on the known inverse relationship between the first-order correlation function and the power spectrum. (They are a Fourier transform pair.) With random motion the frequencies of the scattered light spread about the incident light, producing a Lorentzian distribution (the center being at the frequency of the incident light; the spectrum is known as a Rayleigh line to distinguish it from other spectral

lines such as Raman lines). If an external, polarized electric field is applied to the solution, the molecules drift in the fixed direction as well as diffuse. In the Fourier transform this is equivalent to multiplying the transform by the transform of a uniform linear motion, which happens to be a single sine function. The result in QEL is to replace the constantly decaying relaxation curve by a peak about which the relaxation occurs. The amount this peak is displaced from the zero time of the autocorrelation time scale is a measure of the constant velocity (or its period) of the scattering ions, while from the relaxation time the diffusion coefficient and hence the frictional coefficient can be estimated. Since the time required to collect the autocorrelation data following a single electrical pulse can be as low as 1 msec, this means that short repetitive pulses of relatively high voltage can be applied across the solution and synchronized with the autocorrelation of the scattered photons. Often, in this technique platinum electrodes are used to apply the voltage rather than reversible electrodes. Naturally, with these electrodes, electrolysis of the solvent occurs, and this generates a pH gradient between the electrodes. This effect can be reduced if the polarity of the electrodes is reversed between pulses, but more reliable results could be obtained if reversible silver/silver chloride electrodes were used. The fact that reversing the polarity reverses the movement does not affect the analysis, provided that the voltages of the pulses remain constant and movement is across the incident light beam. If the angle between the incident light and scattered light is decreased, the transport vector decreases in magnitude until at full-forward scattering only the diffusion is important. It is thus possible to discriminate between the two effects by measuring the correlation function at a range of angles.

Few results on the use of this procedure for macroions having relative masses of less than a million have been reported. This is because smaller biological macroions have relatively small scattering cross, sections, and a lengthy experiment is required to obtain a statistically significant result. For small molecules the repetitive pulsing has to extend over many minutes, and in this time extraneous effects, such as the accumulation of electrolysis products (if platinum electrodes must be used) or of gas bubbles, distrub the result. The technique is ideal for simultaneously studying diffusion and charge on a macroion, because diffusion can be converted directly to a frictional coefficient, which means that charge can be correctly calculated from the mobility [see Eq. (6)]. Examples of suitable macroions are viruses and particles from biological cells. Another advantage of the heterodyne method is that it is possible to define a "frequency window" for the analysis of a given size of molecule, so that specific ions can be followed even when present in a mixture of smaller or very much larger ions.

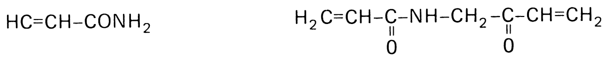
III. ELECTROPHORESIS IN A STATIONARY MATRIX

Despite the decline in the application of free-solution electrophoresis to physical chemistry, the results from this work showed that electrophoresis has a unique place in studies of biological macroions. Thus, attempts were made to reduce the technical problems associated with electrophoresing mixtures of macroions. Efforts were directed at reducing the convective distrubances of the moving boundaries by introducing a neutral but physically inert matrix to support the solutions. Initially, paper saturated with suitable buffers was used, and a spot of the solution of macroion was placed on the paper before the electric field was applied across the strip. The positions of the bands at the end of the experiment were found by selective staining. This was reasonably successful for some applications, but because particles moved in a solvent that was adsorbed to the surface of the cellulose fibers in the paper, there were inconsistencies in the overall conduction of the ions. Heating in regions of low electrical resistance dried the paper unevenly even in a saturated atmosphere.

For these reasons paper was replaced by starch gels, and here the proportion of free solvent to inert matrix was considerably higher than could be achieved with paper. These gels are stabilized by weak intermolecular forces (hydrogen bonding and van der Waals dispersion forces). Thus, they require a backing plate for transferring the gel to a staining medium, and this disturbed the gels during handling. Furthermore, untreated starch has a number of free phosphate groups covalently linked to the carbohydrate chains, and these cause undesirable pumping of water (electroosmosis) when an electric field is applied across the starch strips. These difficulties prompted workers to search for gels that were electrically neutral as well as physically strong enough to be handled without distrotion. This led to the development of modern polyacrylamide gels. These gels form the matrix for most present-day electrophoretic analyses.

A. Polyacrylamide Gel Electrophoresis

Acrylamide has the chemical structure given in Fig. 4a, and it is the opening of the double (π) bond that leads to the polymerization. It is necessary to activate this



(a) (b)

FIGURE 4 Chemical structures of (a) acrylamide and (b) *N,N′*-methylenebisacrylamide (bisacrylamide).

polymerization by one of a variety of procedures, all of which produce free radicals in solution. These radicals are produced by hydrolysis of water, either photochemically using riboflavin phosphate as a catalyst or chemically using chemical catalysts. Once the initiation has occurred, the acrylamide crosslinks with itself, with the evolution of heat to produce linear chains that are terminated when the activated chain ends are neutralized by collision with other free radicals from the solvent. The product is a stable gel containing only a few percentage points (by weight) of polyacrylamide. Gels formed from only acrylamide have many undesirable properties (e.g., they are glutinous and stick to glass), and the pore sizes are illdefined. A more satisfactory gel was formed by including a small proportion of the bifunctional acrylamide (Fig. 4b) in the mixture. This forms random cross-links between the linear chains during polymerization. The average size of the pores formed by this mixture is determined by the proportion of bisacrylamide (Fig. 4b) to the normal acrylamide in the original solution. The mechanical stability of the gel decreases with increased proportion of crosslinking, and for some purposes the bisacrylamide is replaced by N,N'-diallyltartardiamide. The latter produces a gel that is more restrictive to macroions than bisacrylamide at high proportions of cross-linker, but in turn it is more manageable. The solvent is held in these pores, and the whole gel can be handled without mechanical supports, despite the fact that it consists of up to 95% (by weight) of solvent.

Polyacrylamide gel electrophoresis (PAGE) is carried out using simple equipment. The gels are cast either in glass tubes or as slabs supported on nonconducting plates, which can themselves be thermostatted. Because the gels are generally thin, they have low electrical conductance, so relatively high voltages can be applied without excessive heating (say, 1000 V across a 15-cm tube, 2-mm diameter).

Along with the developments of PAGE were the production of a variety of organic ions that would buffer pH without producing solutions of high conductivity. High conductivity arises because of high mobilities of ions, and this is inversely proportional to the van der Waals radius of an ion [Eq. (6); the frictional coefficient is proportional to this radius]. Organic ions have larger radii than inorganic ions such as phosphates, and so their conductivity is less.

B. Molecular Properties Important for Polyacrylamide Gel Electrophoresis

When ions are transported through the pores of the gel, their relative mobilities can be used to determine their sizes, provided that they are not excluded from the pores of the gel. The radii of the transporting ions are not simple to define because for both the stationary matrix and the diffusible ions there are layers of solvent that are transported with the ion (or remain stationary on the gel), therefore making their radii different from that expected from crystallographic models. In general the penetrating ions move through the pores with mobilities commensurate with those measured in free solution. It is usually difficult to measure these mobilities because the path length taken by the ion is unknown, although it is certainly greater than the simple linear distance measured macroscopically along the gel. This means that PAGE cannot yield physicochemical data on absolute charges of diffusible ions; only relative mobilities can be obtained and estimated by reference to standards electrophoresed in the same gel.

In the case of macroions the relationship between relative mobilities and charge is complicated by the possibility that selective retardation of the ions occurs as a result of physical impedence by the gel matrix and adsorption onto the stationary polyacrylamide. Contributions from both occur with all ions, but for proteins the adsorption is less important than the filtering. At the simplest level this filtering would be described through frictional coefficients [Eq. (6)], which for a fixed pore size increase in proportion to the cube root of the mass. Thus, the velocity is proportional to $M^{-1/3}$, and the relative position of macroions on a gel at the end of an experiment are a function of charge and mass (that is, proportional to $Q/M^{1/3}$), which means that the charge cannot be measured independently of the mass. Because of this proportional dependence of relative positions on the two dependent parameters, one must be fixed before the other can be estimated using relative mobilities. Thus, if Q were constant for the macroions in a mixture, the effective van der Waals radii could be estimated for each from its relative mobility, and if the shape of the ions were constant (that is, all spheres), this would give relative masses M.

These criteria have been used extensively for measuring relative masses of proteins and nucleic acids by including in a separate channel a mixture of standard proteins or nucleic acids of known relative masses; the masses are estimated in advance using absolute procedures such as that involving the analytical ultracentrifuge. It is usually not possible to measure charge this way, because a range of macroions are not usually available having a fixed mass but carrying a range of charges. It is possible to generate such a standard by progressive carbamylation of amino groups (easily achieved by cautions warming of a protein with urea solutions) to produce secondary amines that are not charged at neutral pH. If the carbamylation is not allowed to go to completion for all the side chains, this will produce a mixture of molecules of relatively constant mass (the carbamyl group has a relative mass of 60, so even adding 10 groups to one macroion will not significantly increase the mass). This standard mixture can be used to determine relative charges of the unknowns; then if the original absolute charge and amino acid composition

are known (required to give the number of amines substituted), it is possible to calculate the number of charges on the unknown protein. Inaccuracies are introduced into these calculations because the amount of adsorption of the charged macroions to the polyacrylamide is dependent on the charge, so the position of the polyampholyte after carbamylation does not remain a simple function of charge.

This brief definition of the molecular parameters that affect rates of transport in PAGE illustrates the important assumptions made in converting relative positions to molecular parameters for the procedures discussed in the following sections.

C. Estimation of Relative Masses Using Polyacrylamide Gel Electrophoresis

Most of the charge of a biological macroion comes from dissociation of the intrinsic chemical groups. In the case of many of the nonparticulate and soluble proteins, these groups have isoelectric points in the pH range 4–5, which means that at neutral pH they are negatively charged. The absolute charge is not independent of mass, because the capacity to carry more amino acids bearing charged side chains is greater the larger the mass, while the composition is determined by genetic factors. This means the ratio $Q/M^{1/3}$ is not constant for all proteins, so that separations between individual proteins can be obtained experimentally. In the case of nucleic acids the total charge is generally related to mass for a given type of nucleic acid, because here each nucleoside (the effective monomer of nucleic acids) carries free phosphates that are equally ionized at neutral pH. Thus, the relative positions after PAGE can be related to size. As a result, masses determined by a single PAGE experiment with native proteins are less readily interpreted in terms of van der Waals radii than those made with nucleic acids (but see the later discussion of the Ferguson plot). It has been found, however, that when a protein is mixed with certain charged detergents [the most popular being sodium dodecyl sulfate (SDS)] the quantity of detergent associated with a gram of protein is relatively constant. The result of this association is a spheroidal micelle having a charge and frictional coefficient proportional to the relative mass of the protein [see Eq. (6)]. The addition of SDS dissociates multisubunit proteins into their respective components, so although adding SDS produces a macroion whose mass can be estimated from PAGE (the intrinsic charge of the protein is swamped by the added charge from the SDS), the native biologically active units cannot be examined in the detergent. Despite this deficiency, SDS–PAGE has become the most popular method of determining relative masses of protein subunits and has displaced the ultracentrifuge in routine investigations. Another attraction is that the detergent solubilizes otherwise insoluble proteins and peptides, making it possible to study otherwise insoluble mixtures.

When a highly charged macroion is produced, the aggregate moves rapidly in the gel matrix under a moderate potential gradient. Their relative velocities are related to size, but since these macroions still carry some residual shape from the native molecule, the frictional coefficient is not always equal to the expected sphere. Hence, it is important to use reference standards whose overall original shape is similar to the unknown proteins if reasonable estimates are to be obtained. (The absolute accuracy is seldom better than $\pm 10\%$, although reproducibility is much higher.) To reduce the contribution from variable shapes it is usual to perform several experiments in gels formed from various concentrations of acrylamide and bisacrylamide. Plotting the logarithm of the relative mobility against the concentration of acrylamide gives a straight line whose slope can be related to the molecular size, while the intercept on the ordinate (infinite dilution of acrylamide and bisacrylamide) is a measure of the mobility of the SDS–protein in free solution (Ferguson plot). These plots can be used to determine relative masses of native proteins, because the slope is a measure of the effective ratio of charge to mass at unit charge. Proteins that contain a significant amount of covalently linked carbohydrate can still give anomalous results in this plot because the randomly arranged carbohydrate chains change the overall shape of the ellipsoid from that given by standards using purer proteins. Another factor to be considered is the dependence on the amount of detergent bound per unit weight of peptide. Although this is generally constant, there are notable exceptions where the equilibrium between free SDS and that bound does not follow the expected relationship. To overcome this problem, high concentrations of SDS (say, 10% solutions) may be required in some cases, and this has its own limitations.

The bands or spots produced by SDS–PAGE are widened by diffusion of the micelles within the pores of the polyacrylamide, but since the electrophoretic mobilities of the bands are unidirectional and are greater than those produced by diffusion, the leading edge of the band is sharper than that expected from a simple diffusional model [Eq. (9)]. The concentration of the micelles at the leading edge is an advantage when small amounts of a macroion are being studied. Some experimental procedures enhance this sharpening by enlisting the Kohlrausch regulating function [$\mathbf{T}_{\pm}/C$, Eq. (4)]. To produce sharp bands this ratio must be unequal on the two sides of an interface, and to achieve this the salt concentrations (and pH for polyampholytes) must be different across the interface. Practically, this is achieved by layering a thin band of gel containing different buffers on top of the main gel and electrophoresing the protein through this band before entering the main gel. More elaborate arrangements of

different salt concentrations and pH have been used to produce stacking of native proteins by varying their relative charges in a plane in the gel. This changes the relative mobilities, so that $\mathbf{T}_{\pm}/C$ for each protein is not equal to that in the next plane (called isotachophoresis).

Procedures similar to those used for proteins can be used for nucleic acids. Here, the gels not only act as stable supports for mechanical handling, but also as separate mixtures according to mass. No additional detergents are required because the structure of nucleic acids in solution is essentially a random coil formed from highly charged polyelectrolytes.

For technical reasons gels containing less than 2% (by weight) of acrylamide are unmanageable, and yet even at that concentration the pore sizes are too small to admit large macroions (radii >10 nm). To handle these ions, gels must be formed from polysaccharides. The most common is agarose, a polymer of galactose. Agarose is a fraction from agar, a seaweed polysaccharide, which is partly sulfonated so the charged parts must be removed for use in electrophoresis. These gels are mechanically fragile but can be partly stabilized if a few covalent bonds are formed between some of the galactose units.

D. Isoelectric Focusing and Isotachophoresis

The procedures described for PAGE employ the charges only as a means of electrically driving the macroions along the gel in a fixed direction in order to separate mixtures and estimate relative masses. These procedures yield little information on the charge of the protein or utilize the unique pH where polyampholytes have no net charge (see Section II.D) in order to separate a complex mixture. A stationary boundary forms at this pH, but since the object of the experiment is to both separate and concentrate proteins into narrow bands, a stable pH gradient must be generated. To do this the solution containing the monomeric acrylamide must contain ampholyte buffers that when electrophoresed move more rapidly than the polyampholytes and settle at their respective isoelectric points where they buffer the pH. If a wide-ranging mixture of ampholytes, themselves having many different isoelectric points, is used, the result is a stable pH gradient, which can be made approximately linear with distance between the electrodes.

A pH gradient can be formed if water is electrolyzed—acid at the anode and alkaline at the cathode—but the buffering capacity of water is negligible and the gradient is easily swamped by the protein when they are included. To make a stable gradient a variety of organic ampholines have been synthesized, with various proportions of acidic and basic groups in the heterogeneous mixture of ampholines. When these are included in the polyacrylamide gel as free solutes, a pH gradient can be generated by electrophorsing the slab for a period before adding the protein. This pre-electrophoresing settles the ampholines at their varied isoelectric points. Because ampholines contain many functional groups,[1] they possess greater buffering capacity than a similar mass of peptides, and so the gradient is not disturbed by the presence of the protein.

In these experiments the current is very low after the initial removal of excess charged diffusible ions because the only remaining transportable ions in the system are H^+ and OH^-. The focused boundaries follow the approximate shape described by Eq. (21) and are stationary within the gel for long periods. True equilibrium, where the position is independent of very long times, is rarely achieved because the transport of the H^+ and OH^- continues, and this slowly drags the ampholines with them, eventually destroying the gradient. Ampholines have been developed that can be copoymerized into the polyacrylamide gel to prevent their movement, and when these are used the pH gradient is formed mechanically before the acrylamide is polymerized.

With this technique it is possible to focus individual proteins from a mixture into bands that are fractions of a millimeter wide and have effective concentrations exceeding the solubility of the protein. At these high concentrations the protein "steals" the water from the polyacrylamide chains, and this weakens the matrix, making the columns fragile at this point. The technique can be applied to the study of native or denatured proteins. In the latter case a charged detergent (e.g., SDS) is displaced from the protein during electrophoresis because the unassociated detergent moves to the anode, therefore forcing the micelles to dissociate in order to maintain chemical equilibrium between associated and free detergent. Eventually, all the adsorbed SDS is stripped away from the peptide, leaving it with its native intrinsic charge (and possible insolubility).

For preparative procedures where relatively high concentrations of a mixture of polyampholytes are applied initially to the pH gradient, it is often uneconomic to employ supporting ampholines in the gel. In these cases a pH gradient is generated during preparation of the gel column using conventional buffers. The object of an experiment is to isolate one polyampholyte selectively by stacking it at an interface between two zones using the Kohlrausch regulating function as an underlying theoretical guide (see Section I.A). In order to apply the technique it is necessary to have studied the protein using analytical PAGE in order to ascertain its relative charge and mass. The latter is required in order to adjust the properties of the supporting polyacrylamide gel in isotachophoresis so that it does not

[1]They can be likened to polyacrylic acid, where each monomer exposes a charged carboxylic group, but in the ampholines each monomer exposes a basic and an acidic group, so by producing a mixture of oligomers from the ampholines a wide range of isoelectric points can be produced.

restrict the movement of the polyampholyte. Then a series of buffer zones must be inserted, and both the type and the concentrations of these buffers must be controlled so that a pH gradient is generated yet the ionic strength in the stacking regions is low. Then after electrophoresing of the polyampholytes, the selected few that produce equality for the regulating function across the phases produce sharp boundaries at these interfaces. The selection depends on the charge of the polyampholyte (which is dependent on pH in the zone) as well as the buffer concentration. Ideally, the polyampholyte should be the major transporter of the current rather than the buffer ions. Computer programs exist that optimize the buffering conditions for these experiments. The procedure is often known as steady-state stacking and multiphasic zone electrophoresis. Although these procedures have much in common with isotachophoresis, there are experimental variations between them which might distinguish them for some applications.

E. Two-Dimensional Polyacrylamide Gel Electrophoresis

Present applications of electrophoresis are far removed from the early physical work where the technique provided the only means of measuring charges on particles. Now electrophoresis has become an empirical tool for separating complex biological mixtures of macroions. In this evolution, attempts were made to expand to two-dimensional separations, but only recently has this procedure been successful. The method illustrates the resolving power produced by combining isoelectric focusing in one direction with separations according to van der Waals radii in the second. In order to carry out the experiment the protein mixture is first separated according to the isoelectric points of the components (see Section III.D). At the end of this experiment these gels are placed on top of a slab of polyacrylamide, usually formed as a gradient in concentration of acrylamide and increasing in concentration away from the isoelectrically focused gel. This combination is then electrophoresed by SDS–PAGE (Section II.C) before the separated peptides in the gel are visualized by selective staining or autoradiogaphy. Combining the two orthogonal properties of proteins produces the typical result shown in Fig. 5 for a mixture taken from a rat's liver. The procedure does not lend itself to inclusion of reference standards in the gel, while prior denaturation of the proteins with SDS yields a valuable map of the gene products, as in Fig. 5. It is helpful in genetic studies if the disulfide cross-links between the subunits have been destroyed by adding mercapto compunds to the original preparations. These maps are reproducible provided that the experimental procedures are rigorously controlled, making it possible to

FIGURE 5 Two-dimensional separation of a mixture of denatured peptides from a rat's liver. The separation horizontally is by isoelectric focusing (Section III.D) with the low pH to the left. The vertical separation was made by SDS–PAGE according to size, the largest at the top. (The upper limit of relative mass is $\sim 10^5$ and the lower limit is $\sim 15{,}000$.)

locate unique peptides affected by biological experiments through direct superposition of many maps on top of one another. As many as 2000 peptides have been resolved in one map from a single mixture, and individual variations can be noted by suitable computer analyses of the digitized maps.

Two-dimensional maps of nucleic acid fragments are also made during sequencing procedure in order to read gene codes. These maps are different in format from those prepared from protein mixtures because the radioactively labeled and partially hydrolyzed nucleic acids are separated in one direction according to size, while the second dimension contains parallel "ladders" formed from replicate experiments in which the nucleic acids are hydrolyzed by different enzymes. The result is a series of bands that can be correlated with other bands in order to produce the correct sequence of nucleotides in the original nucleic acid. These gels are usually much longer than those used for protein separations, and the visualization is generally carried out after labeling with radioactive isotopes or covalently linked fluorescent dyes.

F. Other Techniques

In general the matrix methods utilize only two properties of ions in their separation, namely, charge and/or size. It is in the visualization of the separated products that the various methods differ. Some procedures employ immunological reactions to visualize the products. Either this is done after the separation by transferring the proteins to other media prior to reacting with antibodies, or the antibodies are included in the acrylamide gel so the reactions occur during the separation. Other methods rely on specific enzyme reactions that produce colored bands in order to locate the enzymes; they cannot be performed on denatured products. Finally, fluorescent labeling before and after electrophoresis is superceding radioactive labeling for simple detection, which requires the development of new physical detecters for scanning the gels.

IV. ELECTROPHORESIS IN OSCILLATING ELECTRIC FIELDS

Electrophoresis is generally associated with transport in a stationary electric gradient, but the velocity of small ions is sufficient for them to have been trasported during half a cycle of an alternating field (see Section II.B). As the frequency of the field is increased, this directional transport of the whole molecule in the solution is replaced by rotation of the polarized molecule until eventually at frequencies above 1 GHz all the mass transport ceases. Above this frequency the motions of groups and individual atoms become increasingly important, and these are studied by spectroscopic procedures. At lower frequencies procedures have been devised to measure the rotational diffusion of polar molecules; among these may be mentioned dielectrophoresis, which unlike electrophoresis is concerned with the motion of both charged and uncharged particles in a nonuniform electric field. The dielectric losses are measured as the frequency of the field is increased, and when the frequency exceeds that where the rotating dipolar molecules can follow the oscillating field there is a marked change in the dielectric properties of the solution. This frequency can be related to the distance between the poles in the dipole, provided that assumptions are made about the shape of the molecule and values for the local viscosity of the solvent can be obtained. In general these observations have not been applied widely because in solutions containing mixture of small and large ions difficulties arise in interpreting the uncertain contributions from relaxations of gegenions or counterions with respect to the macroion. In other words, the intention of the experiments is frustrated by the lack of reliable interpretation. The work must be supplemented by data from other sources in order to describe the electrical properties of shielded dipoles.

SEE ALSO THE FOLLOWING ARTICLES

CAPILLARY ZONE ELECTROPHORESIS • CHEMICAL THERMODYNAMICS • ELECTROCHEMISTRY • ELECTROLYTE SOLUTIONS, TRANSPORT PROPERTIES • GAS CHROMATOGRAPHY • MICELLES

BIBLIOGRAPHY

Foret, F., Krivanková, L., and Boc, P. (1993). "Capillary Zone Electrophoresis," Wiley, New York.

Jandik, P., and Bonn, G. (1993). "Capillary Electrophoresis of Small Molecules and Ions," VCH Publishers, Weinheim/New York.

Khaledi, M. G., ed. (1998). "High-Performance Capillary Electrophoresis: Theory, Techniques, and Applications," Wiley, New York.

Lunn, G. (2000). "Capillary Electrophoresis Methods for Pharmaceutical Analysis," Wiley, New York.

Rabilloud, T. (1999). "Proteome Research: Two-Dimensional Gel Electrophoresis and Detection Methods," Principles and Practice Series, Springer-Verlag, Berlin/New York.

Rothe, G. M. (1994). "Electrophoresis of Enzymes Laboratory Methods," Springer Laboratory Series, Springer-Verlag, Berlin/New York.

Tietz, D., ed. (1998). "Nucleic Acid Electrophoresis," Springer Laboratory Series, Springer-Verlag, Berlin/New York.

Wehr, T., and Rodriguez-Diaz, R. (1998). "Capillary Electrophoresis of Proteins," Dekker, New York.

Weinberger, R. (2000). "Practical Capillary Electrophoresis," Academic Press, San Diego.

Westermeier, R. (1997). "Electrophoresis in Practice: A Guide to Methods and Applications of DNA and Protein Separations," 2nd ed., Wiley, New York.

Elemental Analysis, Organic Compounds

T. S. Ma
City University of New York

I. Introduction
II. Determination of Carbon, Hydrogen, and Nitrogen
III. Determination of Oxygen
IV. Determination of Sulfur
V. Determination of Chlorine, Bromine, and Iodine
VI. Determination of Fluorine
VII. Determination of Arsenic and Phosphorus
VIII. Determination of Metallic Elements
IX. Other Methods for Determining Atomic Ratios

GLOSSARY

Absorption tube Tube that contains reagent to absorb certain gaseous reaction products.

CHN analyzer Apparatus that determines carbon, hydrogen, and nitrogen simultaneously.

CHNS/O analyzer Apparatus that can be used to determine carbon, hydrogen, nitrogen, and sulfur simultaneously, and oxygen separately.

Combustion Heating at high temperatures.

Combustion train Apparatus connected in series for combusting the sample and collecting the reaction products.

Decomposition Breaking down of the organic molecule.

Determination of an element Quantitative analysis to find out the content of a particular element in the sample.

Digestion Heating the sample in a solution.

Fusion Heating the sample mixed with solid reagents.

Kjeldahl flask Flask with pear-shaped bottom and long neck for the determination of nitrogen, phosphorus, and so forth by digestion in strong acids.

Mode of finish Method used to measure the product obtained after the decomposition of the organic compound.

Schöniger flask Conical flask fitted with a stopper sealed to platinum gauze, for use in the determination of sulfur, halogens, and so forth by the closed-flask combustion technique.

ELEMENTAL ANALYSIS for organic compounds is a branch of quantitative chemical analysis. It deals with the principles and methods for determining the percentages of certain chemical elements present in a pure organic compound or in an organic mixture.

I. INTRODUCTION

Every organic compound contains the element carbon together with one or more of the other elements in the periodic table, in definite proportions. Elemental analysis serves to measure the proportion of each element present in the compound. Thus, when a new compound is synthesized in the laboratory or when an unknown compound is obtained from nature, the product is isolated and carefully purified. Milligram amounts of the pure sample are subjected to elemental analysis. The percentage of every element is separately determined if necessary, and the sum of all elements should total 100%. In this way the composition of the synthetic compound is confirmed, while an unknown compound is characterized by the quantitative relationship of its constituents and its empirical formula can be deduced.

Besides being used to determine the elements in pure organic compounds, elemental analysis is performed on mixtures of organic matter to determine the content of certain elements therein. For instance, foodstuffs are analyzed for nitrogen, which gives an indication of the nutritional value of the material. Shale is analyzed for carbon and hydrogen to estimate its fossil oil content. Coal is analyzed for sulfur and nitrogen because these two elements are responsible for environmental pollution due to the burning of coal.

The general principle of elemental analysis of organic compounds involves the total breakdown (decomposition) of the organic molecules so that the individual elements are converted into their respective elemental forms or simple inorganic compounds. So that the objective of quantitative analysis can be accomplished, two conditions must be met: (1) the decomposition of the sample should be complete and (2) the element to be determined should be transformed into one specific product that can be measured accurately.

At present, it is possible to determine any element that may be present in organic materials. There are a great variety of techniques for decomposing the organic sample. After decomposition, there may be more than one method of measuring the product. The latter process is known as the mode of finish. In the following sections, the elements most commonly determined in organic compounds are discussed, and selected methods for decomposition and finishing are described.

II. DETERMINATION OF CARBON, HYDROGEN, AND NITROGEN

A. Determination of Carbon and Hydrogen

With few exceptions, organic compounds always contain both carbon and hydrogen. When an organic compound is decomposed by heating at high temperatures (combustion) in the presence of oxygen, carbon dioxide and water are produced:

$$\text{Organic compound} \xrightarrow{[O]} CO_2 + H_2O. \tag{1}$$

The resultant water vapor and carbon dioxide can be collected sequentially in separate receivers (absorption tubes). From the weights of carbon dioxide and water obtained, the percentages of carbon and hydrogen are calculated by the following formulas:

$$\%C = \frac{\text{wt of } CO_2}{\text{wt of sample}} \times \frac{\text{at wt of C}}{\text{mol wt of } CO_2} \times 100 \tag{2}$$

and

$$\%H = \frac{\text{wt of } H_2O}{\text{wt of sample}} \times \frac{\text{at wt of H}}{\text{mol wt of } H_2O} \times 100. \tag{3}$$

This is the absolute method for the determination of carbon and hydrogen. The apparatus and technique were developed in the early part of the nineteenth century. Then at the turn of the twentieth century, Pregl (Nobel Prize laureate, 1923) improved the method and demonstrated that carbon, hydrogen, and some other elements in organic compounds can be determined accurately by using a few milligrams of the sample.

B. Dumas Method for Nitrogen Determination

If a nitrogen-containing organic compound is decomposed by oxidation, nitrogen oxides are formed:

$$\text{Organic nitrogen compound} \xrightarrow{\text{oxidizing agent}} \text{Nitrogen oxides.} \tag{4}$$

All nitrogen oxides can be converted to elementary nitrogen by passing them through metallic copper at high temperature:

$$\text{Nitrogen oxides} \xrightarrow{\text{Cu}} N_2. \tag{5}$$

The nitrogen gas is then purified, collected, and measured.

C. Simultaneous Determination of Carbon, Hydrogen, and Nitrogen

Currently the common practice in analyzing organic compounds for carbon, hydrogen, and nitrogen is to use the

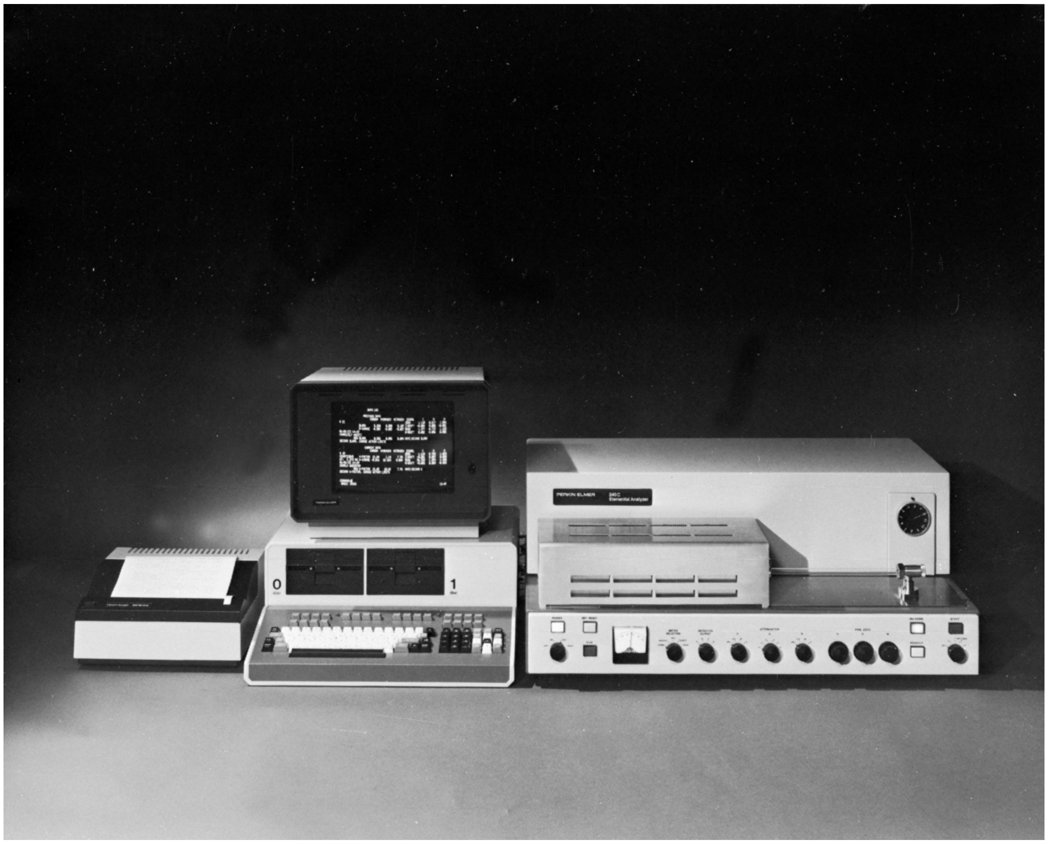


FIGURE 1 Perkin-Elmer 240C Elemental Analyzer connected to a microcomputer. [Courtesy of Perkin-Elmer Corporation.]

CHN analyzer, an apparatus that determines all three elements simultaneously on the basis of the chemical reactions discussed above. Several models of this apparatus are available commercially. An example is shown in Fig. 1, in which the CHN analyzer is connected to a microcomputer. With this equipment, after the accurately weighed sample has been placed in the combustion tube and the sample weight entered on the keyboard, the analysis is started and all operations of combustion: measurement of CO_2, and H_2O and N_2; calculations; and printout of results are carried out automatically. Figure 2 shows a schematic diagram of the combustion train. The boat that contains the organic sample is put in the ladle and introduced into the combustion tube by opening the sample fitting plug. By means of the magnet, the boat is pushed to the high heat coil area. The automated process now commences. The organic compound is combusted in the pyrolysis tube in an oxygen atmosphere under static conditions. The carbon dioxide, water vapor, and nitrogen oxides produced are then swept by a stream of helium into the reduction tube, where nitrogen oxides are converted to nitrogen gas and the excess oxygen is removed. The gaseous mixture is conducted into the analyzer, where the measurement of CO_2, H_2O, and N_2 is performed by three pairs of thermal conductivity cells connected in series. A trap between the first pair of cells absorbs H_2O from the gas mixture before it enters the second cell so that the signal is proportional to the amount of H_2O removed; another trap between the second pair of cells removes CO_2 so that the signal is proportional to the CO_2 removed; and the last pair determines N_2 by comparing the remaining sample gas plus helium with pure helium. The instrument is calibrated with a pure known organic nitrogen compound such as acetanilide before samples are run.

Adaptation of the CHN analyzer for total automatic operation when a large number of samples have to be analyzed within a short period involves changing the combustion train from horizontal to the vertical arrangement. The head of the combustion tube is connected to a turntable device known as the autosampler. Figure 3 depicts the schematic flow diagram of the PE 2400 CHN Analyzer equipped with an autosampler which holds 60 samples in the carousel sample tray. Figure 4 shows a complete picture of this instrument connected to the microprocessor and microbalance. Instead of being placed in an open microboat, each sample is weighed in a tin or an aluminum

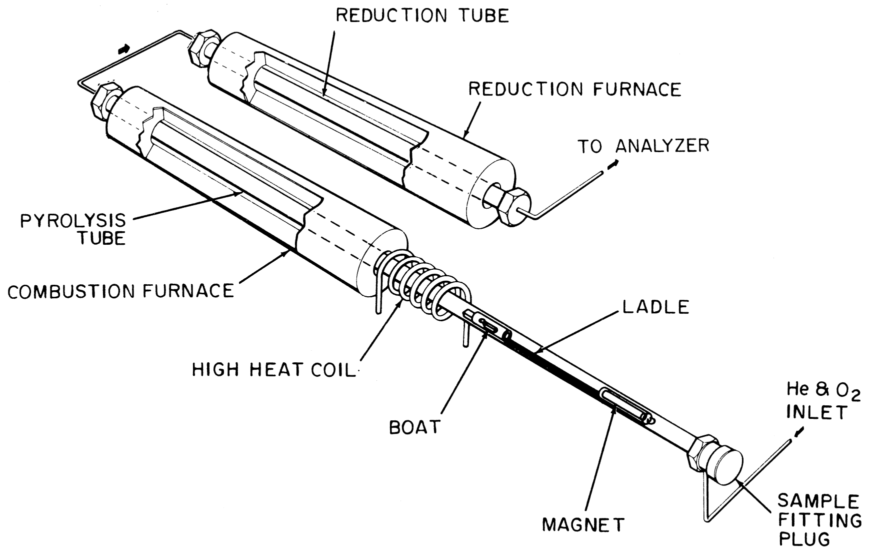


FIGURE 2 Schematic diagram of the combustion train in the Perkin-Elmer 240C Elemental Analyzer. [Courtesy of Perkin-Elmer Corporation.]

capsule. The capsule is then sealed and placed in its corresponding location in the 60-sample tray. Sample weight and identification are stored in memory. As the automatic analysis mechanism is activated, the capsules drop sequentially from the turntable into the combustion tube. The reaction products N_2, CO_2, and H_2O are separated by frontal chromatography and measured by thermal conductivity, respectively. Oxygen is used during the combustion, while helium serves as the carrier gas for chromatography. At the completion of each sample analysis, results are printed out to record the identification number, sample weight, and C, H, and N percentages.

In the Leco CHN Analyzer, infrared spectroscopy is utilized to determine carbon and nitrogen. After combustion, the gaseous mixture containing nitrogen, carbon dioxide, and water vapor is conducted through two infrared absorption cells where H_2O and CO_2 are measured, respectively. Helium serves as the carrier gas.

D. Kjeldahl Method for Nitrogen Determination

Kjeldahl discovered in 1883 that when agricultural materials were heated with concentrated sulfuric acid, the nitrogen originally present in the organic matter was transformed into ammonium bisulfate. He accomplished the determination of nitrogen by liberation of ammonia through the action of a strong alkali and titration of the ammonia. The chemistry involved can be depicted as follows:

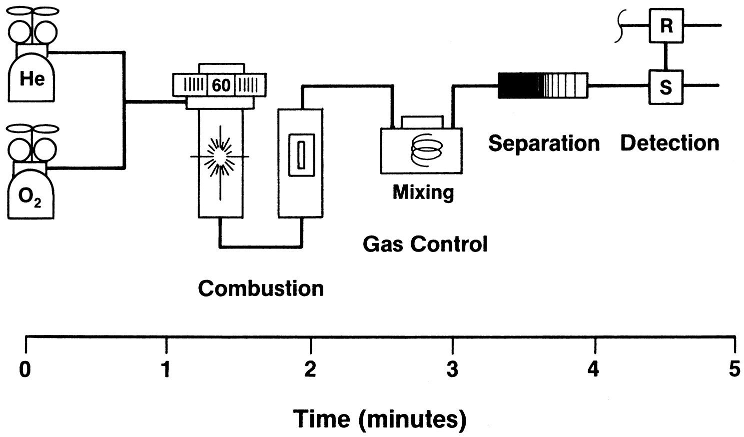


FIGURE 3 Schematic flow diagram of the PE 2400 CHN Analyzer. [Courtesy of Perkin-Elmer Corporation.]

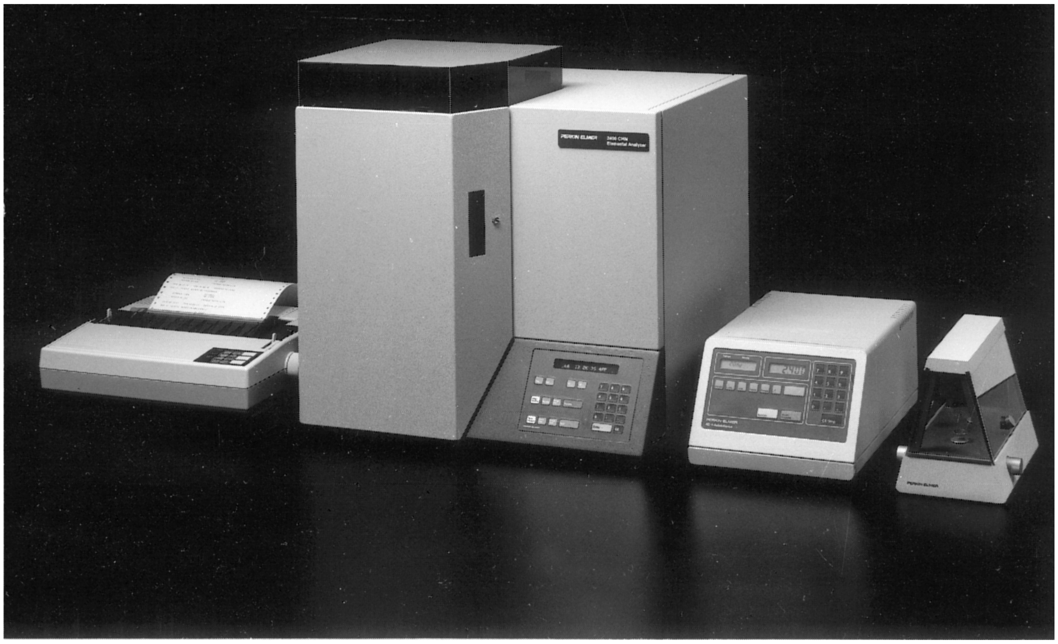


FIGURE 4 The complete assembly of the PE 2400 CHN Analyzer. [Courtesy of Perkin-Elmer Corporation.]

$$\text{Organic nitrogen compound} \xrightarrow{H_2SO_4} (NH_4)HSO_4, \quad (6)$$

$$(NH_4)HSO_4 \xrightarrow{NaOH} NH_3, \quad (7)$$

$$NH_3 \xrightarrow{H_2O} NH_4OH, \quad (8)$$

and

$$NH_4OH + HCl \longrightarrow NH_4Cl + H_2O. \quad (9)$$

The Kjeldahl method is probably the most frequently used method in organic elemental analysis. It is performed generally on complex mixtures, and it is routinely carried out in agricultural stations, food processing plants, and clinical and biochemical laboratories. The equipment and experimental procedures vary widely, depending on the nature of the organic material. Thus, in some determinations milligram quantities of the sample are used, while in other cases the sample size may be as large as 5 g. The reaction vessels employed for the decomposition of the sample range from 10 to 800 mL in capacity. The modes of finish can be titrimetric, colorimetric, or based on the ammonium ion-specific electrode.

A simple procedure for analyzing milligram amounts of organic nitrogen compounds can be carried out as follows. The sample is weighed into a micro-Kjeldahl digestion flask, which is commercially available or can be homemade from a 150-mm test tube by blowing out its bottom to form a bulb of about 10-mL capacity. Ten milligrams of selenium powder and 40 mg of copper sulfate–potassium sulfate mixture are added, followed by 1 mL of concentrated sulfuric acid. The reaction mixture is boiled gently for about 10 min until it becomes colorless. On cooling, the solution is diluted with water and the ammonia is liberated by using the micro-Kjeldahl distillation apparatus shown in Fig. 5. In operation, the ammonium bisulfate solution in the micro-Kjeldahl digestion flask is transferred through funnel B to the bottom of distilling flask D, followed by 8 mL of 30% sodium hydroxide solution. Funnel B is then closed by putting the Teflon plug in place. Steam is then conducted from generator A into flask D, whereupon ammonia is driven from flask D into condenser C. The distillate (ammonium hydroxide solution) is collected in a 50-mL conical flask containing 5 mL of 2% boric acid solution. The ammonium hydroxide is titrated with 0.01 *N* hydrochloric acid [see Eq. (9)], with methyl red–bromcresol green as the indicator.

When the Kjeldahl method is used for nitrogen determination, it should be remembered that the nitrogen present in the organic compound must be the amino type. Other types such as nitro and nitroso compounds can be reduced to amino compounds by suitable treatment prior to concentrated sulfuric acid digestion. Complex organic materials such as coal and blood samples require the addition of catalysts and prolonged heating to achieve complete recovery of nitrogen as ammonium bisulfate.

In the mid-1990s, Collins, Chalk, and Kingston developed a microwave digestion method which eliminates the need for a catalyst and reduces the amount of sulfuric acid

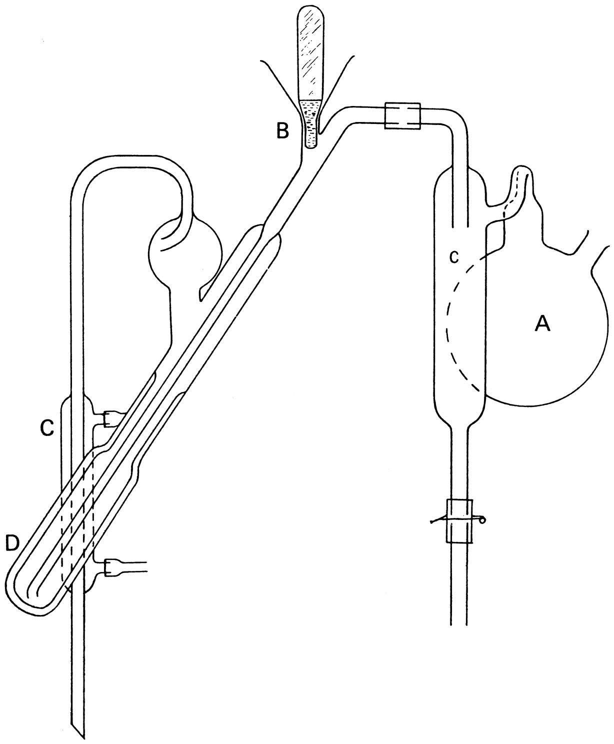


FIGURE 5 Simple micro-Kjeldahl distillation apparatus. [Reprinted with permission from Cheronis, N. D., and Ma, T. S. (1964). "Organic Functional Group Analysis." Wiley, New York Copyright 1964 Wiley.]

considerably. For biological materials, 10 mL of concentrated sulfuric acid per gram sample is recommeneded, followed by 6–12 mL of 30% hydrogen peroxide. The digestion time is about 20 min.

III. DETERMINATION OF OXYGEN

When an oxygen-containing compound is pyrolyzed at 950°C in the presence of platinized carbon in a nitrogen or helium atmosphere, the organic molecule is broken up and all oxygen is converted to carbon monoxide according to the following reactions:

Compound containing C, H, O

$$\xrightarrow{\text{platinized carbon}} CO + CO_2 + H_2O, \qquad (10)$$

$$H_2O + C \longrightarrow CO + H_2, \qquad (11)$$

and

$$CO_2 + C \longrightarrow 2CO. \qquad (12)$$

Hence measurement of the amount of carbon monoxide produced will indicate the oxygen content of the organic substance. In practice, the carbon monoxide is determined indirectly by transforming it to carbon dioxide. This can be achieved by passing the gas through copper oxide at 600°C,

$$CO + CuO \rightarrow CO_2 + Cu, \qquad (13)$$

or through anhydroiodic acid at 130°C,

$$CO + I_2O_5 \rightarrow CO_2 + I_2. \qquad (14)$$

Several finishing modes are applicable. The carbon dioxide can be retained in an absorption tube packed with Ascarite (sodium hydroxide mixed with asbestos) and weighed, or absorbed in a solution and titrated. The iodine produced in Eq. (14) can also be determined by titration. Alternatively, the iodine vapor can be led by a stream of nitrogen into an electrolysis cell, where iodine is reduced at controlled potential and the amount of electricity is recorded. Figure 6 shows, from right to left, the complete equipment for oxygen determination, which comprises the combustion train, the furnace for the oxidation of carbon monoxide, and the assembly for electrometric finish.

The PE 2400 CHN Analyzer (Fig. 4) can be modified to perform automatic determination of oxygen. The combustion tube is filled with platinized carbon reagent. The samples are pyrolyzed in an inert atmosphere of argon or helium. The reaction product CO is separated by frontal chromatography and measured by thermal conductivity. In the Leco CHN Analyzer, carbon monoxide is converted to carbon dioxide to be measured by infrared absorption. Neither CHN method is suitable for the analysis of organic substances which contain fluorine, phosphorus, silicon, or most metallic elements.

IV. DETERMINATION OF SULFUR

Sulfur in organic compounds is usually determined in the form of sulfate. When organic material is heated in a large excess of oxygen, as in the burning of petroleum, its sulfur content is converted to sulfur trioxide, which combines with water to produce sulfuric acid:

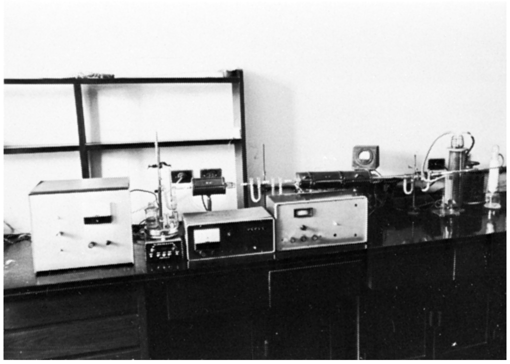


FIGURE 6 Equipment for the determination of oxygen in organic compounds. [Courtesy of Prof. Z. Y. Hu.]

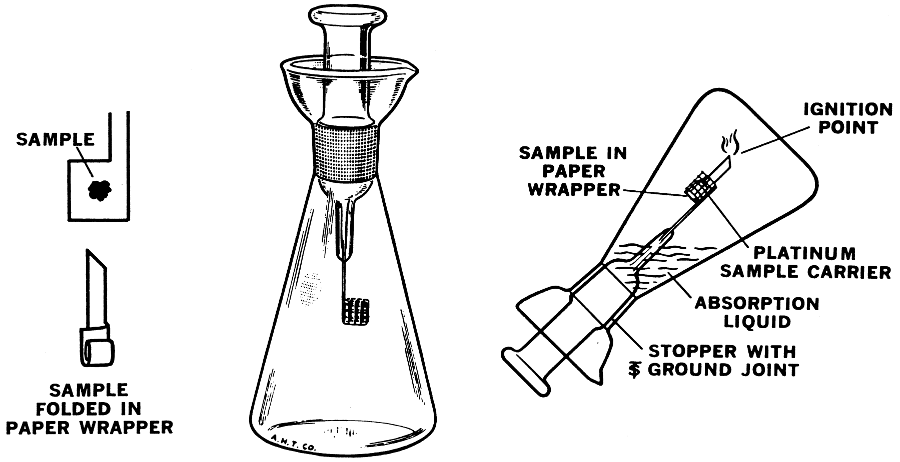


FIGURE 7 Closed-flask combustion technique. [Courtesy of Thomas Scientific.]

$$\text{Organic sulfur compound} \xrightarrow{O_2} SO_3 \qquad (15)$$

and

$$SO_3 + H_2O \rightarrow H_2SO. \qquad (16)$$

A simple technique for decomposing organic substances in a flask filled with pure oxygen is closed-flask combustion (or oxygen-filled flask combustion) with platinum as the catalyst. This method is suitable for sulfur determination. The appartus is illustrated in Fig. 7. When the sample size is between 3 and 25 mg, the flask (Schöniger flask) has a capacity of 500 mL. For the decomposition of larger quantities of organic materials, 1000-mL flasks are used. The stopper of the flask is sealed to platinum gauze. In operation, the sample to be analyzed is placed on a piece of paper cut in the shape shown at the left in Fig. 7. The wide part of the paper is then folded and inserted into the platinum gauze. Meanwhile, 10 mL of distilled water and 0.3 mL of 30% hydrogen peroxide are added to the flask to serve as absorption liquid, and the air inside the flask is displaced with pure oxygen. Next the tip of the paper is ignited and the stopper is immediately attached to the flask. To prevent escape of sulfur trioxide, the analyst must tilt the flask as illustrated at the right in Fig. 7.

After cooling to room temperature, the contents of the conical flask are quantitatively transferred to a 200-mL graduated beaker. The pH of the solution is adjusted to 3.0 ± 0.2 by addition of 0.5 N ammonium hydroxide. Then 50 mL of acetone and 0.3 mL of dimethylsulfonazo-III indicator are added. While the mixture is stirred vigorously, the sulfate ions are titrated with 0.01 M barium chloride:

$$SO_4^{2-} + Ba(ClO_4)_2 = BaSO_4 + 2ClO_4^-. \qquad (17)$$

The end point is a sky-blue color that persists for at least 30 sec.

When the organic compound to be analyzed contains only carbon, hydrogen, oxygen, and sulfur after closed-flask combustion as described above, the resultant solution will be dilute sulfuric acid. In this case, it is more convenient to use acidimetry as the mode of finish. For this purpose, the solution in the Schöniger flask is transferred into a 100-mL conical flask, boiled for 2 min to remove residual hydrogen peroxide, and then titrated with 0.01 N sodium hydroxide, with methyl red as the indicator:

$$H_2SO_4 + 2NaOH = Na_2SO_4 + 2H_2O. \qquad (18)$$

In the CHNS/O analyzer, sulfur is determined in the form of SO_2. The flow diagram of the Leco Analyzer for CHNS/O (Fig. 8) is shown in Fig. 9. The organic sample is weighed in a tin capsule and dropped into

FIGURE 8 Leco Analyzer for CHNS/O. [Courtesy of Leco Corporation.]

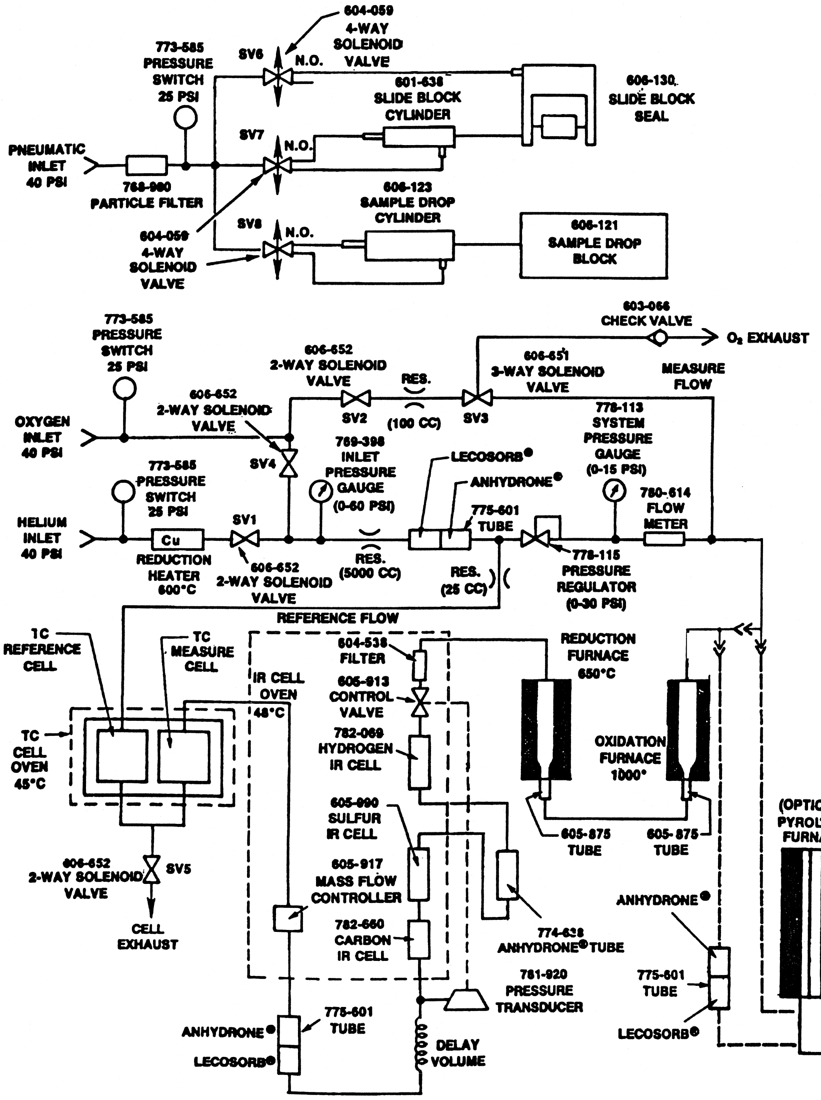


FIGURE 9 Flow diagram of the Leco Analyzer for CHNS/O. The pyrolysis furnace is for oxygen determination. [Courtesy of Leco Corporation.]

the oxidation furnace containing copper oxide and silver tungstate. Heating at 1000°C in an oxygen atmosphere produces CO_2, H_2O, N_2, nitrogen oxides, and sulfur oxides. The gaseous mixture then passes through the reduction furnace containing copper metal at 600°C in a helium atmosphere, whereupon all nitrogen oxides are converted to N_2 and sulfure oxides to SO_2. With helium as the carrier gas, the mixture is conducted through three infrared (IR) cells to measure H_2O, SO_2, and CO_2, sequentially. Finally, these three components are removed by Anhydrone (magnesium perchlorate) and Lecosorb (sodium hydroxide), which leaves N_2 to be measured by thermal conductivity. The Perkin-Elmer Analyzer uses gas chromatography to separate all four components. Neither CHNS/O method is applicable to the analysis of materials containing metallic elements which form thermally stable sulfates.

A special sulfur test apparatus is used for the determination of sulfur in petroleum products by lamp combustion. The apparatus consists of the burner flask, chimney, spray, and absorption tube. The sulfate that is collected can be determined by a turbidimetric method.

V. DETERMINATION OF CHLORINE, BROMINE, AND IODINE

A. Decomposition Methods

The nature of the organic substance containing chlorine, bromine, or iodine determines what is the best method for decomposition. On the one hand, organic compounds that have ionizable halogens, such as the alkaloid halides,

are soluble in aqueous solution, which liberates the halide ions, which can be titrated directly. Similarly, some compounds can be dissolved in a nonaqueous solvent like ethyl alcohol or liquid ammonia; then addition of metallic sodium dislodges the halogen from the organic molecule and produces sodium halide. On the other hand, certain compounds, especially the polyhalogenated ones, require drastic reactions at high temperatures to destroy the organic molecule in order to convert the halogens into ionic forms. The closed-flask combustion technique (see Fig. 7) is commonly used for this purpose. The absorption liquid for chlorine or bromine contains sodium hydroxide and hydrogen peroxide and that for iodine contains hydrazine sulfate, so that chloride, bromide, and iodide, respectively, are obtained as the final products.

Another technique for decomposing organic halogen compounds utilizes a metal bomb constructed of nickel. Commercial metal bombs are available in two sizes: a 2.5-mL bomb for decomposing up to 50 mg of organic material, and a 22-mL bomb that can handle as much as 0.5 g of sample mixed with 15 g of solid reagents. The reagents are sodium peroxide and sucrose or potassium nitrate. After being locked tightly, the bomb is heated. Vigorous oxidation reactions take place, resulting in the transformation of the chlorine, bromine, and iodine originally present in the organic substance to chloride, bromide, and iodate, respectively.

B. Modes of Finish

After decomposition, the resultant chloride and bromide can be determined gravimetrically by weighing the respective silver halides:

$$Cl^- + AgNO_3 \rightarrow AgCl + NO_3^- \quad (19)$$

and

$$Br^- + AgNO_3 \rightarrow AgBr + NO_3^-. \quad (20)$$

This mode of finish is recommended for occasional runs of one or two samples. For a series of determinations or routine analysis, the expedient method is to titrate the solution containing chloride or bromide with standardized silver nitrate solution and locate the end point potentiometrically.

In the Mitsubishi organic halogen measurement system, chloro and bromo compounds in the environment (air, water) are adsorbed in columns filled with activated carbon and then heated at 900°C in a horizontal combustion tube in an oxygen atmosphere. The hydrogen chloride and hydrogen bromide produced are introduced into the automatic titration cell and determined coulometrically with silver ions.

For the determination of iodine, one mode of finish consists of titration with standardized mercuric nitrate solution, with diphenylcarbazone as the indicator:

$$2I^- + Hg(NO_3)_2 \rightarrow HgI_2 + 2NO_3^-. \quad (21)$$

Another method suitable for the determination of small amounts of iodine in organic materials involves the amplification technique. The iodide ions obtained after decomposition of the sample are oxidized to iodate by the addition of bromine in an acetate buffer. Excess bromine is removed with formic acid. Then the iodate is determined by liberation of iodine on addition of iodide in sulfuric acid solution, followed by titration of the liberated iodine with standardized sodium thiosulfate solution with starch as the indicator. The sequence of reactions can be depicted as follows:

$$I^- + 3Br_2 + 3H_2O \xrightarrow{\text{acetate buffer}} HIO_3 + Br^- + 5HBr, \quad (22)$$

$$\text{Excess } Br_2 + HCOOH \longrightarrow 2HBr + CO_2, \quad (23)$$

$$HIO_3 + 5KI + 5H_2SO_4 \longrightarrow 3I_2 + 3H_2O + 5KHSO_4, \quad (24)$$

and

$$3I_2 + 6Na_2S_2O_3 \longrightarrow 6NaI + 3Na_2S_4O_6. \quad (25)$$

VI. DETERMINATION OF FLUORINE

Although fluorine belongs to the halogen group in the periodic table, the methods described above for the determination of chlorine, bromine, and iodine are not suitable for the determination of fluorine. For the decomposition of organic compounds containing fluorine, the method of choice is fusion with metallic sodium or potassium in a metal bomb. The drastic reducing action converts organically bound fluorine to alkali fluoride. So that all possible interfering substances can be removed prior to the mode of finish, fluoride is recovered as fluorosilicic acid from the fusion mixture by steam distillation in perchloric acid solution maintained at 135°C:

$$6F^- + SiO_2 + 6HClO_4 \rightarrow H_2SiF_6 + 6ClO_4^- + 2H_2O. \quad (26)$$

Figure 10 shows a series of three distillation apparatuses used for routine analysis. Figure 11 describes the construction of the distilling head and demonstrates the technique of steam distillation from a solution at a constant temperature above 100°C.

As to the finishing mode, the resultant fluoride can be determined by titrimetry with thorium nitrate or by

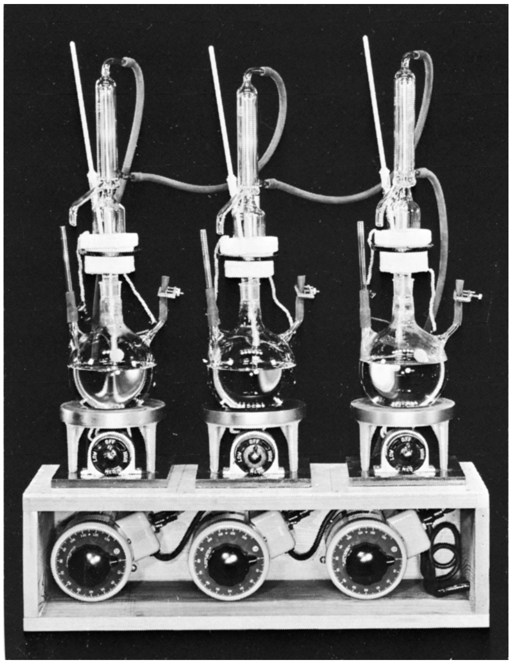


FIGURE 10 Series of three distillation apparatuses for the determination of fluorine. [Reprinted with permission from Ma, T. S., and Rittner, R. C. (1979). "Modern Organic Elemental Analysis." Dekker, New York. Copyright 1979 Dekker.]

spectrophotometry based on the formation of colored complexes such as those produced by reaction with zirconium-SPANS [sodium-2-(p-sulfophenylazo)-1, 8-dihydroxynaphthalene-3,6-disulfonate], cerium alizarin, and zirconium erochrome cyanine. The potentiometric method with the fluoride ion-selective electrode can also be used.

VII. DETERMINATION OF ARSENIC AND PHOSPHORUS

Organic compounds containing phosphorus or arsenic are decomposed by heating (digestion) with a mixture of nitric and sulfuric acids, which forms phosphate and arsenate ions, respectively:

$$\text{Organic phosphorus compound} \xrightarrow{HNO_3, H_2SO_4} PO_4^{3-} \quad (27)$$

and

$$\text{Organic arsenic compound} \xrightarrow{HNO_3, H_2SO_4} AsO_4^{3-}. \quad (28)$$

After digestion, the phosphate obtained is usually determined colorimetrically. When the phosphorus content of the organic substance is higher than 5%, it is preferable to use the yellow color produced by the addition of ammonium vanadate and ammonium molybdate to the phosphate solution. If the amount of phosphate in the digestion mixture is below the milligram level, the blue color produced by the addition of ammonium molybdate followed by a reducing agent is recommended. Because the colored solution contains complexes of more than one species, the

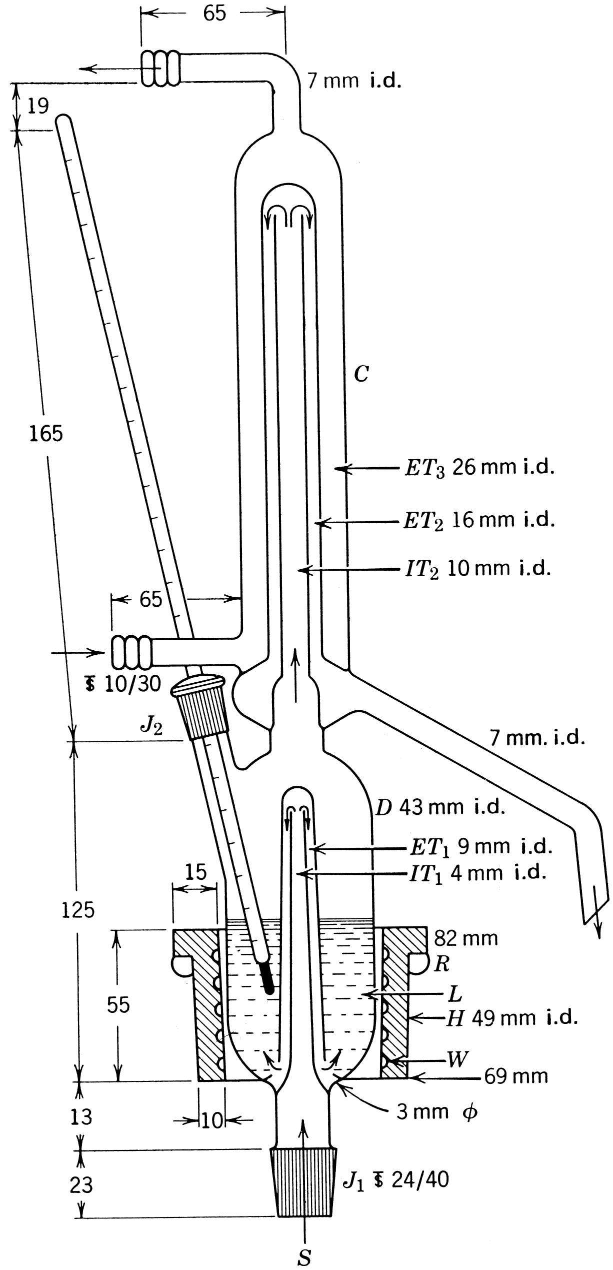


FIGURE 11 Distilling head for steam distillation from a solution at a constant temperature above 100°C. [Reprinted with Permission from Ma, T. S., and Gwirtsman, J. (1957). *Anal. Chem.* **29,** 141. Copyright 1957 Analytical Chemistry.]

experimental procedure should be strictly controlled in order to get reproducible results.

Arsenate can be determined by an iodimetric method in which it is reduced by iodide in acidic solution to produce iodine:

$$AsO_4^{3-} + 2H^+ + 2I^- \rightarrow I_2 + AsO_3^{3-} + H_2O. \quad (29)$$

The iodine liberated is then titrated with a standardized sodium thiosulfate solution [see Eq. (25)].

If the organic sample contains a low percentage of arsenic, arsenate in the digestion mixture is preferably reduced by means of sodium borohydride to liberate arsine, which is measured by atomic absorption spectrometry.

VIII. DETERMINATION OF METALLIC ELEMENTS

A. Determination by Ashing

When a metallic element is a constituent of an organic compound, it will leave a solid residue on complete decomposition of the sample by strong heating. This provides a simple technique, known as ashing, for determining metallic elements, except for mercury and osmium, which form volatile products. The organic substance can be heated in a crucible or microboat: decomposition by heat is aided by moistening the sample with nitric or sulfuric acid. After heating, noble metals such as platinum leave the pure metal:

$$\text{Organic platinum compound} \xrightarrow{HNO_3} Pt. \quad (30)$$

Other metallic elements are converted to oxides, for example,

$$\text{Organic copper compound} \xrightarrow{HNO_3} CuO, \quad (31)$$

or are recovered in the form of sulfates, for instance,

$$\text{Organic sodium compound} \xrightarrow{H_2SO_4} Na_2SO_4. \quad (32)$$

If the organic sample contains a single metallic element, the residue is weighed and its metal content can be calculated. If two or more metallic elements are present in the original substance, the residue is brought into solution and subjected to inorganic analysis.

B. Determination of Mercury

Because of its volatility, mercury in organic compounds cannot be determined by the ashing technique described above. Instead, the organic sample is placed in a porcelain microboat and heated in a combustion tube packed with granulated calcium oxide. The mercury vapor that forms is conducted to a tube containing gold foil, which traps the mercury to produce amalgam. The increase in weight of the gold tube gives the content of mercury in the sample.

Amalgamation in a gold trap is also employed in environmental analysis of organic mercury compounds such as methyl mercury. The mercury is commonly determined by cold vapor atomic absorption spectroscopy.

When an organic mercury compound is decomposed by digestion in a mixture containing nitric acid, sulfuric acid, and potassium permanganate, the resultant mercuric ions are preferably determined by titration with standardized potassium thiocyanate solution:

$$\text{Organic mercury compound} \xrightarrow{H_2SO_4, HNO_3, KMnO_4} Hg^{2+} \quad (33)$$

and

$$Hg^{2+} + 2KCNS \longrightarrow Hg(CNS)_2 + 2K^+. \quad (34)$$

Ferric ammonium sulfate is used as the indicator, and the end point is located spectrophotometrically.

IX. OTHER METHODS FOR DETERMINING ATOMIC RATIOS

In the methods described in the previous sections, a known weight of an organic compound is taken for analysis. From the yield of the product related to a certain element, the weight percentage of that element in the sample is determined. when two or more elements have been determined, their atomic ratio in the organic compound can be deduced on the basis of their atomic weights. Besides the above methods, there are other analytical techniques for determining atomic ratios without weighing the samples. Some are discussed below.

When an organic compound containing hydrogen, oxygen, and nitrogen is pyrolyzed in a large excess of sulfur in the absence of air, carbon is converted to carbon dioxide, carbonyl sulfide, and carbon disulfide; hydrogen forms hydrogen sulfide; oxygen produces carbon dioxide and carbonyl sulfide; and nitrogen is liberated as nitrogen gas:

$$\text{CHON compound} \xrightarrow{S} CO_2 + COS + CS_2 + H_2S + N_2. \quad (35)$$

Using these reactions, Hara *et al.* constructed an apparatus for determining the atomic ratios of carbon, hydrogen, oxygen, and nitrogen in an organic compound. A quartz ampoule of 5.5-mm inside diameter, 6.5-mm outside diameter, and 50-mm length is prepared. Purified sulfur is

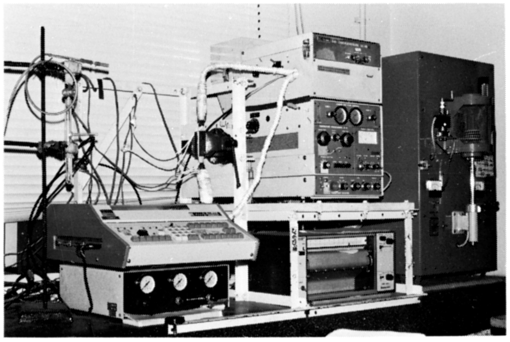


FIGURE 12 Equipment for the determination of carbon, hydrogen, oxygen, and nitrogen ratios by the sulfurization technique. [Courtesy of Prof. T. Hara.]

placed at the bottom of the ampoule, followed by the organic sample. Air inside the ampoule is displaced by a stream of helium with a special displacement device (see Fig. 12), and the ampoule is quickly sealed. The ampoule is then heated at 1500°C in the high-frequency induction furnace for 1 min and left there for 3 min. Subsequently, the ampoule containing the reaction products is inserted into a Teflon sampler, which is connected to the gas chromatograph. The ampoule is broken by driving a jack across the Teflon sampler, and the mixture of gaseous products is introduced into the gas chromatograph, where N_2, CO_2, H_2S, COS, H_2O, SO_2, and CS_2 are separated and their peak areas are recorded. Correction factors are obtained by analyzing five standard compounds. The atomic ratios of carbon, hydrogen, oxygen, and nitrogen in an unknown compound are determined by using these correction factors. Figure 12 shows the complete equipment: the air-displacement device is on the left side; the Teflon sampler and crushing device are at the center, with the connecting tubes to the gas chromatograph; and the high-frequency induction furnace is on the right side of the assembly.

It is obvious that the technique of Hara *et al.* described above is not applicable to the analysis of gaseous samples and low-boiling liquids. For volatile organic compounds, Yu *et al.* put forth a device that uses high-resolution glass capillary gas chromatography in combination with microwave plasma emission spectrometry. The method is based on the complete destruction of a compound into its constituent atoms by means of the helium plasma, followed by measurement of each atomic species in the spectrometer. The signal obtained for a specific element is nearly proportional to the quantity of the element; thus it is possible to calculate the elemental ratios of the compound according to the formula

$$\text{Elemental ratio} = \frac{\text{elemental ratio of standard} \times \text{element signal ratio of unknown}}{\text{element signal ratio of standard}}. \quad (36)$$

The reference standard is optional. The apparatus comprises three parts: gas chromatograph, microwave generator and plasma discharge tube, and polychromator and data recording. The gas chromatograph serves to separate mixtures of organic compounds and submit pure samples for decomposition. The spectrometer has channels for carbon, hydrogen, oxygen, nitrogen, fluorine, chlorine, bromine, iodine, phosphorus, sulfur, and mercury. This method is particularly suited for the analysis of petroleum products and halogenated hydrocarbons.

Hughes, Brown, and Fry described the photodiode array application of the red and near-infrared region of the argon inductively coupled plasma atomic emission spectrum from 650 to 900 nm for nonmetallic elements deriving from the atomization of pure organic compounds. It was found that C, H, N, and O lines of reasonable intensity can be obtained with short exposure times, and the red and near-infrared spectra of these elements are unusually simple, with no apparent line interferences and very low levels of plasma continuum background. Hence it is feasible to use this technique for the simultaneous determination of carbon, hydrogen, nitrogen, and oxygen in the organic sample.

When the atomic ratios of all elements present in an organic compound have been determined, the empirical formula of the compound can be calculated. The empirical formula can be checked against the molecular weight found by a suitable method. Alternatively, the chemical formula of an organic compound can be obtained with the aid of high-resolution mass spectrometry. Since the latter technique can measure molecular species to a precision of seven digits, it is possible to deduce the chemical formula by matching the molecular weight (mass) with the various combinations of atoms of the elements expected to be present in the compound. It is prudent, however, to verify the chemical formula by determining one or two elements by chemical methods. Obviously, mass spectrometry is not applicable to the determination of any element in a mixture of organic compounds.

SEE ALSO THE FOLLOWING ARTICLES

Combustion • Gas Chromatography • Organic Chemistry, Compound Detection • Organic Chemistry, Synthesis

BIBLIOGRAPHY

Belcher, R., ed. (1977). "Instrumental Organic Elemental Analysis," Academic Press, London.

Collins, L. W., Chalk, S. J., and Kingston, H. M. S. (1996). "Atmospheric pressure microwave sample preparation procedure for the combined

analysis of total phosphorus and Kjeldahl nitrogen," *Anal. Chem.* **68,** 2610.

Hara, T. *et al*. (1977, 1978, 1980, 1981, 1982, 1983). "Simultaneous determination of the atomic ratio between C, H, O, and N by the pyrolytic sulfurization method," *Bull. Chem. Soc. Jpn.* **50,** 2292; **51,** 1110, 2951, 3079; **53,** 951, 1308; **54,** 2956; **55,** 329, 2127, 3450, 3800; **56,** 1378, 3615.

Hughes, S. K., Brown, R. M., Jr., and Fry, R. C. (1981). "Photodiode array studies of near infrared and red atomic emissions of C, H, N, and O in the argon inductively coupled plasma," *Appl. Spectrosc*. **35,** 396.

Kirsten, W. (1983). "Organic Elemental Analysis," Academic Press, New York.

Ma, T. S. (1997). "Organic elemental analysis," *In* "Analytical Instrumentation Handbook" (G. W. Ewing, ed.), Chap. 3, Marcel Dekker, New York.

Ma, T. S., Gutterson, M., and Wang, C. Y. (1982, 1984, 1986, 1988, 1990). "Fundamental reviews—Organic elemental analysis," *Anal. Chem*. **54,** 87R; **56,** 88R; **58,** 144R; **60,** 175R; **62,** 78R.

Ma, T. S., and Hassan, S. S. M. (1982). "Organic Analysis Using Ion-Selective Electrodes," Vol. 2, Academic Press, London.

Yu, W. L. *et al*. (1981). "Development and application of a prototype $(GC)^2$-MFS hyphenated apparatus," *In* "Proceedings of the 4th International Symposium on Capillary Chromatography" (R. E. Kaiser, ed.), p. 445, Hüthig, Heidelberg.

Gas Chromatography

Milos Novotny
Indiana University

I. Gas Chromatographic Molecular Separation
II. Physical Principles
III. Separation Columns
IV. Detectors and Ancillary Techniques
V. Instrumentation
VI. Applications

GLOSSARY

Mobile phase Gas containing the compounds to be separated while migrating through the stationary phase.

Stationary phase Solid or liquid whose physical properties affect the separation of the mobile-phase compound.

GAS CHROMATOGRAPHY (GC) is a physical separation method used extensively in scientific investigations, chemical practice, petroleum technology, environmental pollution control, and modern biology and medicine. Its primary role is to separate various chemical compounds that were introduced into the system as a mixture and to determine quantitatively their relative proportions. When combined with other physicochemical methods, GC can also provide qualitative (structural) information on the separated substances. The method is limited to relatively volatile (low-molecular-weight) compounds. The principle of separation is a relative affinity of various mixture components to the stationary phase (a solid or a liquid), while the mobile phase (a gas) migrates them through the system. GC is a dynamic separation method, where the separation of components occurs in a heterogeneous phase system.

I. GAS CHROMATOGRAPHIC MOLECULAR SEPARATION

Separating chemical substances from each other has been extremely important to various branches of science and technology for many years. Simple separation procedures such as distillation, crystallization, precipitation, and solvent extraction have been used by humankind from time immemorial. More refined forms of separation, such as chromatography and electrophoresis, have been among the major causes of scientific revolution during the last several decades of this century.

Gas chromatography is one of the several chromatographic methods. The scientific principles of

TABLE I Comparison of Types of Chromatography

Mobile phase	Stationary phase	Types of chromatography[a]	Abbreviation	Separation
Liquid	Solid	Liquid–Solid	LSC[b]	Adsorption
Liquid	Liquid (immiscible)	Liquid–Liquid	LLC	Solubility (partition)
Gas	Solid	Gas–Solid	GSC	Adsorption
Gas	Liquid	Gas–Liquid	GLC	Solubility (partition)

[a] Gas–gas and solid–solid equilibria do occur in nature; however, they are impractical for chromatographic separations.
[b] C = Chromatography.

chromatography were discovered by a Russian botanist M. S. Tswett (1872–1919) but hardly developed into useful chemical separation procedures until the 1930s. The name *chromatography* was originated by Tswett who primarily investigated plant pigments (*chromatos* is the Greek name for color). However, any method that utilizes a distribution of the molecules to be separated between the mobile phase (a gas or a liquid) and the stationary phase (a solid or a liquid that is immiscible with the mobile phase) now qualifies as chromatography. The physical state of the mobile phase determines whether we deal with gas or liquid chromatography.

Variation in the type of stationary phase is important as well: if a solid is used as the stationary phase, the interaction of the molecules under separation with it is due to adsorption forces; if a liquid is used in the same capacity, the molecules under separation interact with it based on their solubilities. According to this type of interaction, we distinguish between adsorption chromatography and partition chromatography. This classification is further evident in Table I.

Tswett's original work pertained to liquid adsorption chromatography, while the first experiments on liquid partition chromatography were described in the early 1940s by A. J. P. Martin and his co-workers in Great Britain (a decade later, recognized by a Nobel Prize in Chemistry). Several investigations pertaining to the use of gas as the mobile phase in gas/adsorption systems were reported in Austria, Czechoslovakia, Russia, and Sweden during the 1940s. However, the development of gas–liquid chromatography, reported in 1952 by A. T. James and A. J. P. Martin, is widely considered the beginning of GC as a powerful analytical method.

Today, GC is complementary to other separation methods. It can be practiced on either a small (analytical) scale or a large (preparative or industrial) scale. The preparative uses of GC are relatively uncommon. While typical amounts of chemical substances analyzed by the modern GC are between the microgram (10^{-6} g) and nanogram (10^{-9} g), samples as small as a femtogram (10^{-15} g) can be measured in special circumstances. Importantly, contemporary GC can often simultaneously recognize up to several hundred chemical substances.

II. PHYSICAL PRINCIPLES

The apparatus designated to separate compounds by GC is called the gas chromatograph. Its essential parts are shown in Fig. 1. At the heart of the system is the separation

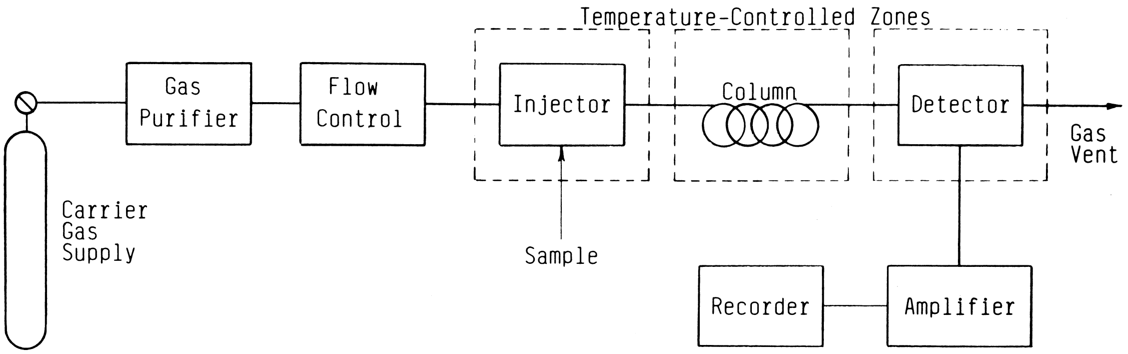


FIGURE 1 A gas chromatograph with its main components.

column, at which the crucial physicochemical processes of the actual compound separation occur.

The separation column contains the stationary phase, while the mobile phase (frequently referred to as the carrier gas) is permitted to flow through this column from a pressurized gas cylinder (source of the mobile phase). The rate of mobile-phase delivery is controlled by a pressure and/or flow-regulating unit. An exclusive separation mode for the analytical GC is elution chromatography, in which the sample (a mixture of chemicals to be separated) is introduced at once, as a sharp concentration impulse (band), into the mobile-phase stream. The unit where sample introduction is performed is called the injector. The unfractionated sample is transferred from the injector into the chromatographic column, where it is subjected to a continuous redistribution between the mobile phase and the stationary phase. Due to their different affinities for the stationary phase, the individual sample components eventually form their own concentration bands, which reach the column's end at different times. A detector is situated at the column's end to sense and quantitatively measure the relative amounts of these sample components.

The detector, together with auxiliary electronic and recording devices, is instrumental in generating the chromatogram, shown in Fig. 2. Such a chromatogram is, basically, a plot of the sample concentration versus time. It represents the individual component bands, separated by the chromatographic column and modified by a variety of physical processes into a peak shape. The position of a peak on the time scale of the total chromatogram bears some qualitative information, since each chromatographic peak represents at least one chemical substance. The areas under the peaks are, however, related to the amounts of individual substances separated in time and space.

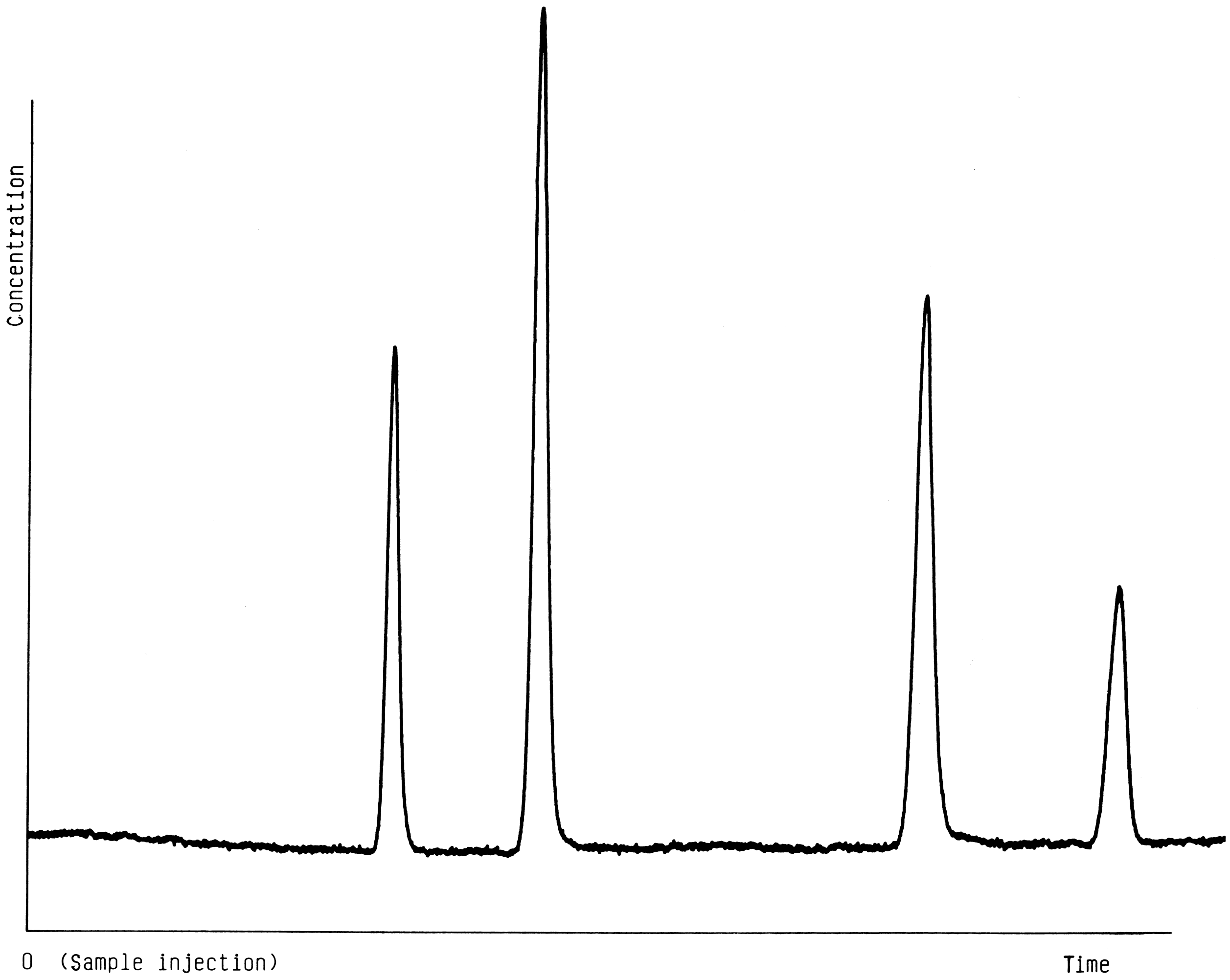


FIGURE 2 Chromatogram.

A typical gas chromatograph has three independently controlled thermal zones: proper temperature of the injector zone ensures rapid volatilization of the introduced sample; the column temperature is controlled to optimize the actual separation process; and the detector must also be at temperatures where the individual sample components are measured in the vapor phase. For certain GC separations, it is advisable to program the temperature of the chromatographic column.

As shown in Fig. 2, different sample components appear at the column's end at different times. The retention time t_R is the time elapsed between injection and the maximum of a chromatographic peak. It is defined as

$$t_R = t_0(1 + k), \quad (1)$$

where t_0 is the retention time of a mixture component that has no interaction with the stationary phase (occasionally referred to as dead time), and k is the capacity factor. The capacity factor is further defined as

$$k = K\frac{V_s}{V_M}, \quad (2)$$

where K is the solute's distribution coefficient (pertaining to a distribution between the stationary phase and the mobile phase), V_s is the volume of the stationary phase, and V_M is the volume of the mobile phase in a chromatographic column. The distribution coefficient $K = C_s/C_M$ (where C_s is the solute concentration in the stationary phase and C_M is the solute concentration in the mobile phase) is a thermodynamic quantity that depends on temperature as do all equilibrium constants. The molecular interactions between the phases and the solutes under separation are strongly temperature-dependent. If, for example, a solid adsorbent (column material) is brought into contact with a permanent (inorganic) gas and a defined concentration of organic (solute) molecules in the gas phase at a certain temperature, some solute molecules become adsorbed on the solid, and others remain in the permanent gas. When we elevate the system temperature, less solute molecules are adsorbed, and more of them join the permanent gas; the distribution (adsorption) coefficient, as defined above, changes correspondingly. Likewise, if the stationary phase happens to be a liquid, the solute's solubility in it decreases with increasing temperature, according to Henry's law, resulting in a decrease of the distribution (partition) coefficient.

According to Eqs. (1) and (2), the retention time in GC depends on several variables: (a) the chemical nature of the phase system and its temperature, as reflected by the distribution coefficient; (b) the ratio of the phase volumes in the column V_s/V_M; and (c) the value of t_0. In the practice of chromatography, these variables are used to maximize the component separation and the speed of analysis.

Unlike some other chromatographic processes, the physical interactions between the mobile phase and solute molecules in GC are, for all practical purposes, negligible. Thus, the carrier gas serves only as means of molecular (solute) transport from the beginning to the end of a chromatographic column. The component separation is then primarily due to the interaction of solute molecules with those of the stationary phase. Since a variety of column materials are available, various molecular intertactions can now be utilized to enhance the component separation. Moreover, these interactions are temperature-dependent.

For the mixture component with no affinity for the stationary phase, the retention time t_0 serves merely as the marker of gas linear velocity u (in cm/s) and is actually defined as

$$t_0 = \frac{L}{u}, \quad (3)$$

where L is the column length. The gas velocity is, in turn, related to the volumetric flow rate F since

$$u = \frac{F}{s}, \quad (4)$$

where s is the column cross-sectional area. The gas-flow rate is chiefly regulated by the inlet pressure value; the higher the inlet pressure the greater the gas-flow rate (and linear velocity) becomes, and consequently, the shorter t_0 is. The retention time t_R of a retained solute is also modified accordingly. Correspondingly, fast GC separations are performed at high gas-inlet pressures. The so-called retention volume V_R is a product of the retention time and volumetric gas-flow rate:

$$V_R = t_R F. \quad (5)$$

Since the retention times are somewhat indicative of the solute's nature, a means of their comparison must be available. Within a given chemical laboratory, the relative retention times (the values relative to an arbitrarily chosen chromatographic peak) are frequently used:

$$\alpha_{2,1} = \frac{t_{R_2}}{t_{R_1}} = \frac{V_{R_2}}{V_{R_1}} = \frac{K_2}{K_1}. \quad (6)$$

This equation is also a straightforward consequence of Eqs. (1) and (2). Because the relative retention represents the ratio of distribution coefficients for two different solutes, it is frequently utilized (for the solutes of selected chemical structures) as a means to judge selectivity of the solute–column interactions.

For interlaboratory comparisons, the retention index appears to provide the best method for documenting the GC properties of any compound. The retention index system compares retention of a given solute (on a logarithmic scale) with the retention characteristics of a set of standard solutes that are the members of a homologous series:

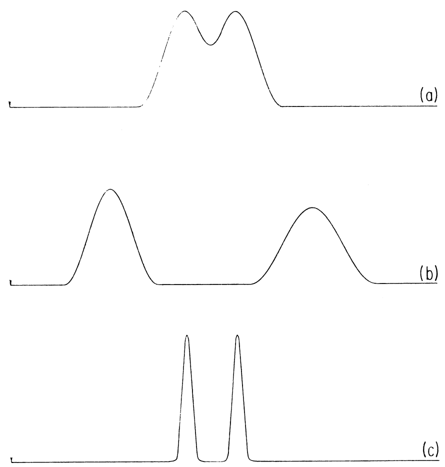


FIGURE 3 Enhancement of component resolution as based on the selectivity and efficiency of the separation process: (a) two unresolved components, (b) resolution based on the column selectivity, and (c) resolution based on the column kinetic efficiency.

$$I = 100z + 100\frac{\log t_{R(x)} - \log t_{R(z)}}{\log t_{R(z+1)} - \log t_{R(z)}}. \tag{7}$$

The subscript z represents the number of carbon atoms within a homologous series, while x relates to the unknown. For example, a series of n-alkanes can be used in this direction; each member of a homologous series (differing in a single methylene group) is assigned an incremental value of 100 (e.g., 100 for methane, 200 for ethane, and 300 for propane, etc.) and if a given solute happens to elute from the column exactly half-way between ethane and propane, its retention index value is 250). Retention indices are relatively independent of the many variables of a chromatographic process.

The success of GC as a separation method is primarily dependent on maximizing the differences in retention times of the individual mixture components. An additional variable of such a separation process is the width of the corresponding chromatographic peak. Whereas the retention times are primarily dependent on the thermodynamic properties of the separaton column, the peak width is largely a function of the efficiency of the solute mass transport from one phase to the other and of the kinetics of sorption and desorption processes. Figure 3 is important to understanding the relative importance of both types of processes.

In Fig. 3, (a) depicts a situation where two sample components are eluted too closely together, so that the resolution of their respective solute zones is incomplete; (b) represents a situation where the two components are resolved from each other through choosing a (chemically) different stationary phase that retains the second component more strongly than the first one; and (c), which shows the same component retention but much narrower chromatographic peaks, thus represents the most "efficient" handling of the two components. This efficiency, represented by narrow chromatographic zones, can actually be attained in GC practice by a proper design in physical dimensions of a chromatographic column. Width of a chromatographic peak is determined by various column processes such as diffusion of solute molecules, their dispersion in flow streamlines of the carrier gas, and the speeds by which these molecules are transferred from one phase to another.

An arbitrary, but the most widely used, criterion of the column efficiency is the number of theoretical plates, N. Figure 4 demonstrates its determination from a

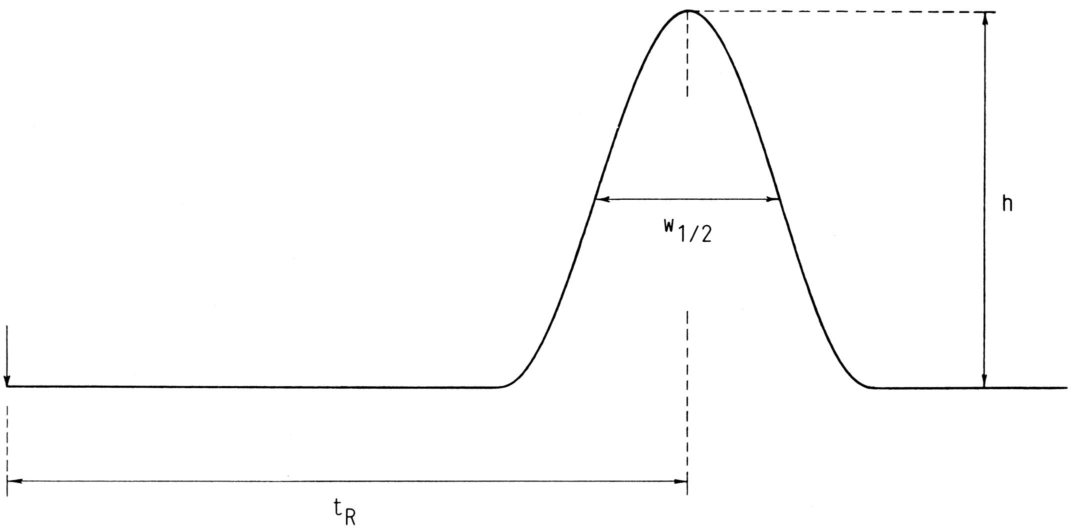


FIGURE 4 Determination of the number of theoretical plates of a chromatographic column.

chromatographic peak. This number is simply calculated from the measured retention distance t_R (in length units) and the peak width at the peak half-height $W_{1/2}$:

$$N = 5.54\left(\frac{t_R}{w_{1/2}}\right)^2. \quad (8)$$

The length of a chromatogrphic column L is viewed as divided into imaginary volume units (plates) in which a complete equilibrium of the solute between the two phases is attained. Obviously, for a given value of t_R, narrower peaks provide greater numbers of theoretical plates than broader peaks. Turning once again to Fig. 3, we see that cases (a) and (b) represent low column efficiencies (plate numbers), while case (c) demonstrates a high-efficiency separation.

Equation (8), used to determine the number of theoretical plates, relates to a perfectly symmetrical peak (Gaussian distribution). While good GC practice results in peaks that are nearly Gaussian, departures from peak symmetry occasionally occur. In Fig. 5, (a) is usually caused by a slow desorption process and undesirable interactions of the solute molecules with the column material, and (b) is associated with the phenomenon of column overloading (if the amount of solute is too large, exceeding saturation of the stationary phase, a fraction of the solute molecules is eluted with a shorter retention time than the average). When feasible, GC should be carried out at the solute concentrations that give a linear distribution between the two phases.

The length element of a chromatographic column occupied by a theoretical plate is the plate height (H):

$$H = \frac{L}{N}. \quad (9)$$

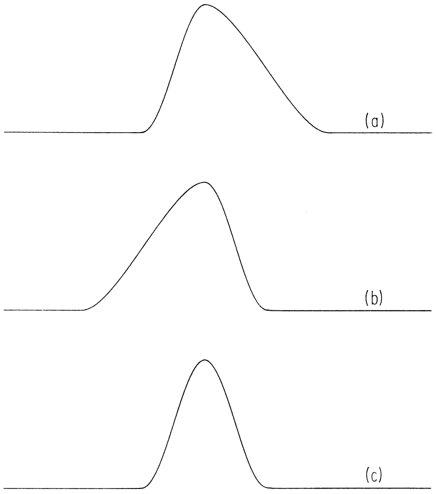


FIGURE 5 Departures from peak symmetry: (a) slow desorption process and (b) column overloading. (c) Gaussian distribution.

The column efficiency N can be dependent on a number of variables. Most importantly, the plate height is shown to be a function of the linear gas velocity u according to the van Deemter equation:

$$H = A + \frac{B}{u} + Cu, \quad (10)$$

where the constant A describes the chromatographic band dispersion caused by the gas-flow irregularities in the column. The B-term represents the peak dispersion due to the diffusion processes occurring longitudinally inside the column, and the C-term is due to a flow-dependent lack of the instantaneous equilibrium of solute molecules between the gas and the stationary phase. The mass transfer between the two phases occurs due to a radial diffusion of the solute molecules.

Equation (10) is represented graphically by a hyperbolic plot, the van Deemter curve, in Fig. 6. The curve shows the existence of an optimum velocity at which a given column exhibits its highest number of theoretical plates. Shapes of the van Deemter curves are further dependent on a number of variables: solute diffusion rates in both phases, column dimensions and various geometrical constants, the phase ratio, and retention times. Highly effective GC separations often depend on thorough understanding and optimization of such variables.

III. SEPARATION COLUMNS

Since the introduction of GC in the early 1950s, many different column types have been developed, as is widely documented by numerous column technology studies reported in the chemical literature. The column design is extremely important to the analytical performance and utility for different sample types and applications. The most important features include (a) type of column sorption material (in both physical and chemical terms), (b) column diameter, (c) column length, and (d) surface characteristics of a column tubing material. A proper combination of these column design features can often be crucial to a particular chemical separation.

Based on their constructional features, GC columns can be divided into three main groups: packed columns, capillary (open tubular) columns, and porous-layer open tubular columns. Their basic geometrical characteristics are shown in Fig. 7.

A packed column is basically a tube, made from glass or metal, that is filled with a granular column material. The material is usually held in place by small plugs of a glass wool situated at each column's end. During a GC run, such a column is attached to the instrument through a gas-tight connection; the carrier gas is forced through the

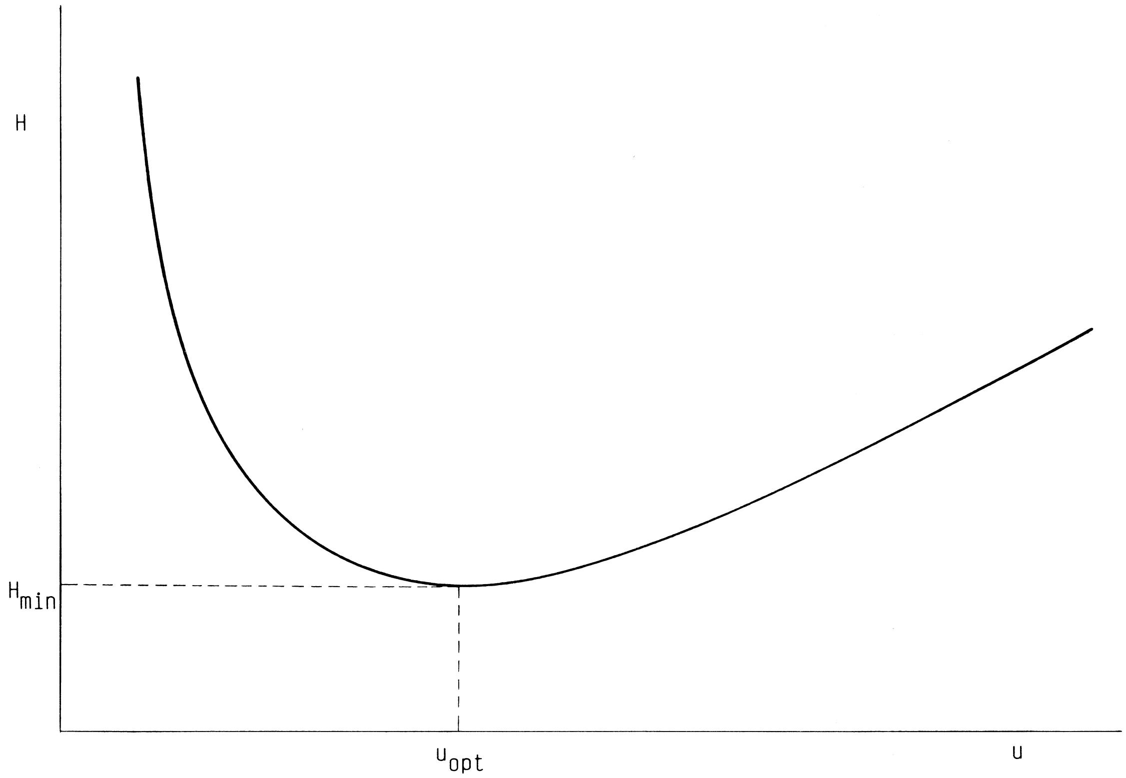


FIGURE 6 Relationship of the plate height and linear gas velocity (van Deemter curve).

free channels between the individual particles, while the sample molecules are allowed to interact with the particles. Typical column inner diameters are 1–4 mm, and the lengths are 1–3 m, although departures from these dimensions may exist for special applications. The inner diameters of preparative columns can be considerably larger.

The granular packing can be either an adsorbent (if the method of choice is gas–solid chromatography) or an inert solid support that is impregnated with a defined amount of a liquid stationary phase (for gas–liquid chromatography). In either case, packing materials with uniformly small particles are sought, as the column performance is strongly dependent on the particle size. In fact, a distinct advantage of small particles is their closer contact with diffusing sample molecules and a greater number of the mentioned equilibrium units (i.e., theoretical plates). Because extremely small particles present a great hindrance to gas flow, materials with a particle size between 100 and 150 μm are typically used as a sensible compromise between the column efficiency and technological limitations of high gas pressure at the column inlet.

In gas–solid chromatography, the solute molecules interact with the surface of solid adsorbents through relatively weak physical adsorption forces. Such weak forces are desirable, because the adsorption process must be reversible, preserving the chemical integrity of the solutes (unlike in some forms of contact catalysis, where a strong compound adsorption precedes chemical conversion). Consequently, not all adsorbing solids qualify as suitable column packings in GC. Examples of suitable GC adsorbents are silica gel, alumina, zeolites, carbonaceous adsorbents, and certain porous organic polymers. Surface porosity and a relatively large surface area are among the characteristic features of GC adsorbents. For example, certain synthetic zeolites, molecular sieves, may have a specific surface area as high as 700–800 m^2/g.

Specificity of certain solute-adsorbent interactions is a major advantage of gas–solid chromatography. Various adsorbents readily discriminate between different molecular geometries of otherwise similar solutes (e.g., geometrical isomers). At present, however, major difficulties exist as well: (1) large distribution coefficients (compared with partitioning liquids) result in long retention times; (2) the separaton process can often be strongly dependent on sample size, which is a serious problem for analytical determinations; (3) the physical processes in adsorption chromatography are less amenable to a rigorous theoretical description compared with gas-partition chromatography; and (4) the current adsorbent technology does not permit an effective suppression of minor catalytic

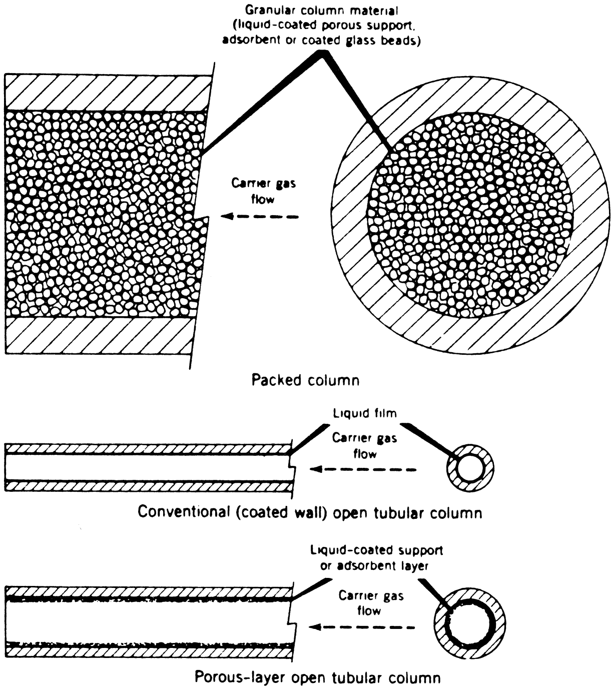
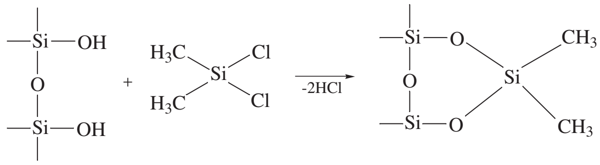


FIGURE 7 The three major types of GC columns. [From Horvath, C. (1967). *In* "The Practice of Gas Chromatography" (L. S. Ettre and A. Zlatkis, eds.), p. 133. Wiley (Interscience), New York.]

effects. The current applications of gas–solid chromatography are largely confined to the separation of relatively small molecules (such as permanent gases and lower alkanes). They represent a relatively small fraction of all GC applications.

Gas–liquid chromatography has found considerable use in chemical analysis. The packing materials (solid supports) utilized in this method are macroscopically similar to the described adsorbents. Yet their function is entirely different: They serve only as a supporting medium for the liquid stationary phase and do not participate directly in the separation process. The specific surface area of such solid supports is considerably less than that of adsorbents (i.e., their microstructure is considerably less developed).

The most commonly used solid supports are the diatomaceous earths. They are fossil-originated minerals found in abundance in various parts of the world. Prior to their use in chromatography, the diatomaceous earths are washed, thermally treated, chemically modified, and sieved, in a large manufacturing process. The diatomaceous earths are basically siliceous materials that contain reactive surface structures, the silanol groups. Since such groups could adversely affect the chromatographic analyses, causing "tailing" of certain polar sample components, they are effectively blocked (deactivated) by a silylation reaction, an example of which is given below:



(a part of original surface structure) (dimethyldichlorosilane. a deactivation agent) (deactivated surface)

The solid support is subsequently impregnated by a liquid stationary phase. While many solid supports can carry up to 25% by weight of a liquid phase before becoming visibly wet, much lower phase loadings (a few percent) are used in practice. Both the amount and the chemical type of a stationary phase are crucial to the separation characteristics (efficiency and sample capacity) of a chromatographic column. Packed columns are considered to be low-efficiency, high-capacity GC columns. While their best efficiencies amount to no more than a few thousand theoretical plates, packed columns can tolerate microgram amounts of samples. Only carefully and totally packed columns yield the expected efficiencies.

The concept of the open tubular (capillary) column was introduced in 1956 by a Swiss scientist, M. J. E. Golay. Due to their extremely high separation efficiencies, open tubular columns have recently revolutionized analytical separations. As seen in Fig. 7, there is no granular packing inside the capillary column. The stationary liquid phase is uniformly deposited as a thin film on the surface of the inner wall, along the entire length of a long column. Typical lengths of capillary columns range from 10–100 m, with 0.2–0.5 mm inner diameters. The columns of smaller diameters (50–100 μm) have also been prepared for extremely efficient separations. Capillary columns with inner diameters larger than 0.5 mm are uncommon.

Column efficiencies between 10^5 and 10^6 theoretical plates have been achieved in capillary GC. Very narrow chromatographic peaks elute from such columns, allowing a high degree of resolution among the individual components of complex mixtures. The resolution advantage of a capillary column over a packed column is clearly indicated by Fig. 8, where numerous constituents of Calmus oil are separated from each other using a capillary column (a). A packed column (b) shows considerably less component resolution. The high separation efficiency of capillary columns is due to their high permeability to the carrier gas (the absence of column packing); consequently, long columns, featuring a great number of theoretical plates, can be prepared.

Another outstanding feature of GC capillary columns is their geometrical simplicity and consequent accessibility

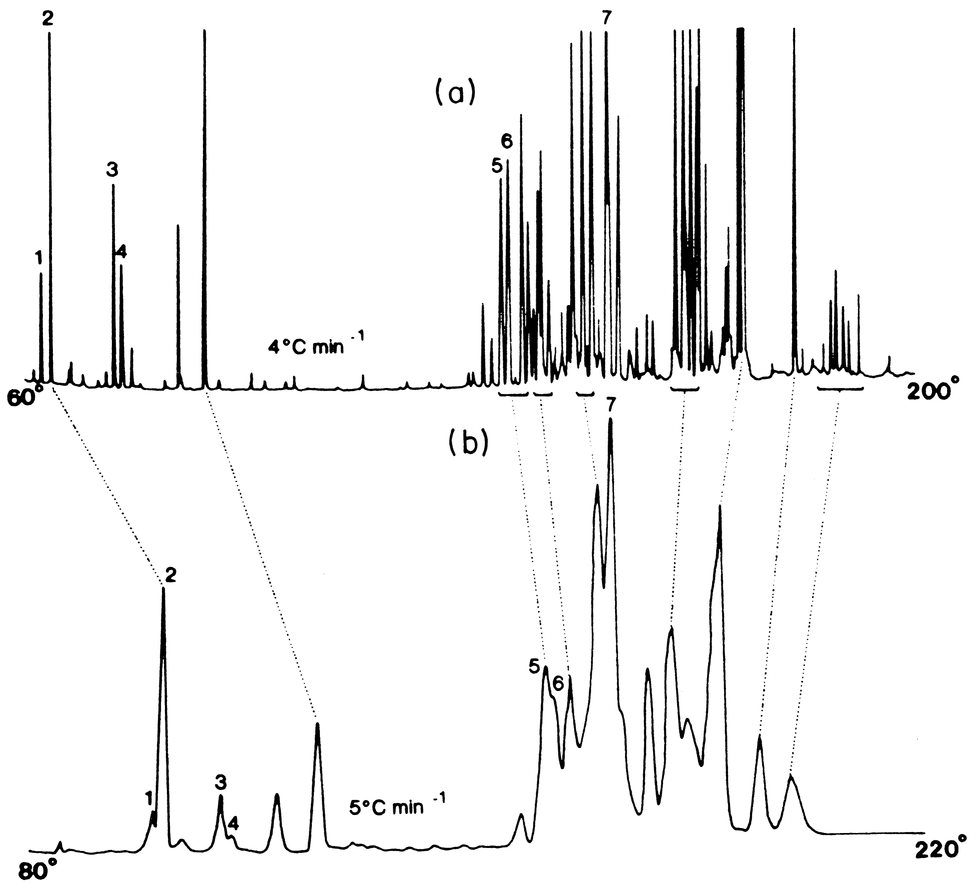


FIGURE 8 The resolution advantage of a capillary column (a) over a packed column (b), in Calmus oil analysis. [From Grob, K., and Grob, G. (1979). *J. High Resolut. Chromatogr.* **3,** 109.]

to theoretical description. An example is a description of the physical processes that occur inside such capillary columns. Equation (10), the van Deemter equation, can be translated, for the capillary column into

$$H = \frac{2D_G}{u} + \frac{(1 + 6k + 11k^2)}{(1 + k)^2} \frac{r^2}{24D_G} u, \qquad (11)$$

where D_G is the solute diffusion coefficient in the mobile (gas) phase and r the capillary inner radius. The equation shows explicitly how the plate height is dependent on the diffusion processes and the column radius. At low gas velocities, molecular diffusion significantly increases the plate height. At higher velocities, the opposite is true (D_G is in the denominator), reflecting the fact that the solute mass transfer from one phase to another is primarily diffusion controlled. Reducing the column radius is a powerful way to increase the column performance.

Note that Eq. (11) is an accurate description of the column processes because the column geometry is well defined. Although somewhat similar equations exist for the packed columns, various (less accurate) empirical constants must be used.

Equation (11) is strictly valid only for the cases where the stationary phase film thickness amounts to no more than a few tenths of a micrometer (thin-film columns). Columns with film thicknesses up to several micrometers can also be prepared. Although their efficiencies are lower than those obtained for the thin-film columns, due to the impaired solute mass transfer, they can tolerate larger sample amounts without signs of overloading.

Refined aspects of column technology have been crucial to the success of GC capillary columns. Early in the development of such columns, metal or plastic tubes were used exclusively. Highly efficient glass capillary columns were developed at a later stage, and the problems of glass fragility were successfully overcome through the technology of fused-silica flexible tubes. Production of fused-silica capillaries is reminiscent of the fabrication of optical fibers: thin-walled sillica tubes, drawn from a hot zone are immediately protected by an overcoat of a stable organic polymer (Fig. 9).

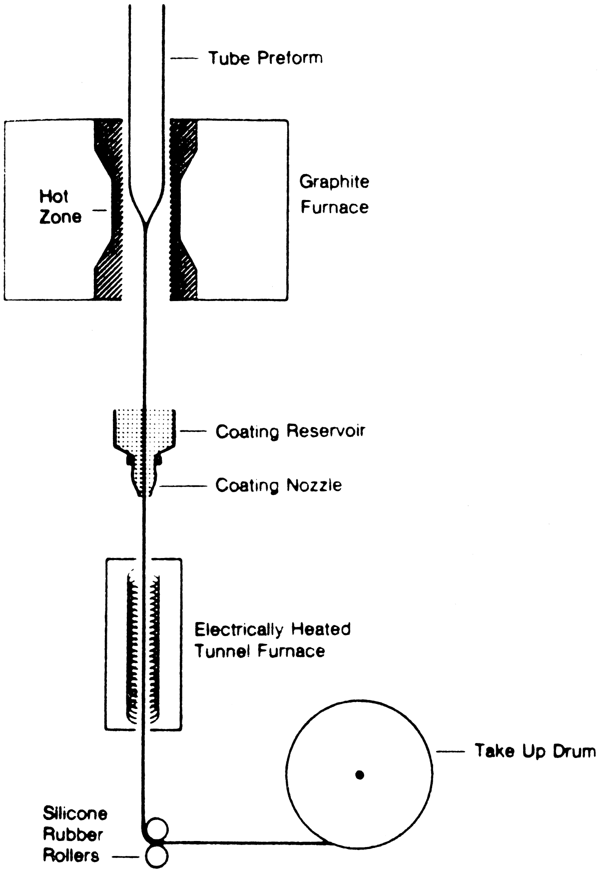


FIGURE 9 Column drawing apparatus for the preparation of fused silica capillaries. [From Lipsky, S. R., McMurray, W. J., Hernandez, M., Purcell, J. E., and Billeb, K. A. (1980). *J. Chromatogr. Sci.* **18,** 1.]

To secure a uniform film deposition from the solution of a stationary phase, the inner column's surface is first treated by an organic compound, the task of which is to improve the surface wettability and to mask potentially adsorptive column sites. A variety of stationary phases can now be successfully coated for capillary GC with a controlled film thickness. Refined procedures now exist even for the chemical immobilization of some stationary liquids.

The third type of a GC column (Fig. 7) is a porous-layer, open tubular column. While such a column has an inner diameter and a length comparable to the wall-coated columns, its inner wall is modified through a chemical treatment or deposition of finely dispersed particles. The porous layer can be either an adsorbent or a thin layer of the solid support impregnated with a liquid stationary phase. Efficiencies of the porous-layer, open tubular columns are not as high as those of "true" capillary columns, but their greater sample capacity is an advantage to some separations. The enhanced sample capacity is obtained because of a greater surface area of such columns and, consequently, a relatively larger amount of stationary phase.

A proper choice of the liquid stationary phase is exceedingly important to a successful chromatographic separation. A great number of chemically different stationary phases have been described in the scientific and commercial literature. Several requirements govern the choice of a chemical substance as a GC stationary phase. First of all, it should have adequate selectivity for the substances to be resolved. It must be chemically stable at the column temperatures used in a given separation problem. The stationary phase must easily adhere as a uniform film to the column support without running off the column; if such mechanical instability occurs, the phase contaminates the detector and, naturally, the columns function properly for only a limited time. Finally, the stationary phase should be a well-defined chemical compound, so that the column preparation as well as the chromatographic process itself are reproducible.

In spite of the above strict requirements, many chemical substances can adequately perform as the stationary phase. The thermal stability requirement has made various synthetic polymers (silicones, polyglycols, polyesters, polyimides, etc.) most popular. Since the stabilities vary according to chemical stucture, nonpolar polymers are more stable than the polar column substrates. Column temperatures above 300°C are seldom used in the practice of GC.

The general solution rules roughly determine the suitability of a stationary phase for a given separation task: Polar substances (solutes) are readily dissolved and chromatographically retained by the polar stationary phases, while the nonpolar column materials retain the chromatographed sample components according to their boiling points, without any particular regard to the presence of unique functional groups in the sample molecules. Although the rules appear relatively straightforward, the stationary phases for many practical separations are still selected empirically.

The main solute–column interactions can be classified as dispersion forces and dipole–dipole interactions. The dispersion forces are present in any solute–solvent system, a hydrocarbon solute interacting with a nonpolar paraffin being often shown as an example. The polar solute molecules have permanent dipoles that can interact with those of the polar phases; on occasions, the dipole moments can also be induced in certain solute molecules in the presence of highly polar column materials. Dipole–dipole interactions are clearly evident in the separations of alcohols, esters, amines, aldehydes, and so on, on the polyglycol, polyamide, polyester, or cyanoalkylsilicone stationary phases.

Some extremely selective GC separations have been accomplished. In a number of cases, the hydrogen-bonding

mechanism has been utilized. Synthesized optically active polymers are highly effective in resolving various racemic mixtures. Certain metal chelates, used as additives to the common stationary phases, can retain selected solutes through the formation of reversible complexes. Finally, highly organized liquids (such as various liquid crystals) tend to retain more strongly the molecules of elongated rather than bulky structures.

Through advances in synthetic chemistry and polymer research, new GC stationary phases will become available. Additional column selectivities can also be achieved by mixing the existing stationary phases with each other, in suitable proportions.

IV. DETECTORS AND ANCILLARY TECHNIQUES

The detector has an extremely important role in the overall process of GC analysis. The current popularity and success of GC as an analytical method is attributable in great part to the early development of highly sensitive and reliable means of detection. In sensing the vapor concentration at the column outlet, the detector provides information on the distribution of individual peaks within a chromatogram (which compound?) as well as their relative amounts (how much?). The area measured under a chromatographic peak is generally related to the quality of the compound.

Many detection principles in GC were investigated over the years, but only a few pass the criteria of reliability needed for precise analytical measurements. Detectors can broadly be classified as universal or selective. Universal detectors measure all (or nearly all) components of a mixture, although their response to the same quantities of different compounds is seldom similar. Selective detectors respond only to mixture components that possess a unique structural feature in their molecules. For example, a typical gasoline sample contains a number of organic components which, after being separated by an appropriate chromatographic column, are all detected by a universal detector. However, if a lead-selective detector is used instead, only a few peaks are recorded, those due to the lead-containing additives in gasoline, while the remaining mixture constituents are ignored. The so-called ancillary techniques go a step further as highly selective detectors, because they actually characterize the individual GC peaks qualitatively.

A. DETECTORS

The most important analytical properties of a GC detector are sensitivity, linearity over an extensive concentration range, long-term stability, and ease of operation. While most GC determinations are performed with solute quantities between 10^{-6} and 10^{-9}, certain selective detectors can reach down to the 10^{-15}-g levels, representing some of the most sensitive measurement techniques available to the chemist. Some GC detection principles are based on the measurement of certain transport properties of the solutes (e.g., thermal conductivity or optical properties), while other detectors are transducers, measuring ultimately some product of a solute molecule (e.g., gas-phase ionization products). The latter detectors are destructive to the solutes.

1. The Thermal-Conductivity Detector

This detector, occasionally referred to as the hot-wire detector or katharometer, operates on the basis of measuring the difference in thermal conductivity of pure carrier gas and the carrier gas plus a solute. Typically, the column effluent is passed through a thermostatted cavity (measuring cell) that contains a resistor element heated by passage of a constant current. Various changes in the thermal conductivity of the surrounding gas causes the element temperature (and, consequently, its electric resistance) to decrease or increase. Pure carrier gas is passed, under the same conditions, through a reference cell of identical design. The resistor elements of both cells are parts of a Wheatstone bridge circuit that records any imbalance caused by the passage of individual solutes.

In thermal-conductivity measurements, it is advisable to choose a carrier gas that differs maximally from the organic solutes (e.g., hydrogen or helium). The detector is a truly universal and simple device, but its sensitivity is marginal; at best, submicrogram amounts are detected. The thermal-conductivity detector is most typically employed for the analysis of permanent gases and light hydrocarbons.

2. The Flame-Ionization Detector

This detector is the workhorse of GC. It operates on the basis of decomposing the solute-neutral molecules in a flame into charged species and electrically measuring the resultant changes of conductivity. A cross-sectional view of a flame-ionization detector is shown in Fig. 10. A small flame is sustained at the jet tip by a steady stream of pure hydrogen, while the necessary air (oxidant) is supplied through the diffuser. At the detector base, the column effluent is continuously introduced, mixed with hydrogen, and passed into the flame. Conductivity changes between the electrodes are monitored, electronically amplified, and recorded. A conventional carrier gas contributes little to the flame conductivity; however, when organic solute molecules enter the flame, they are rapidly ionized,

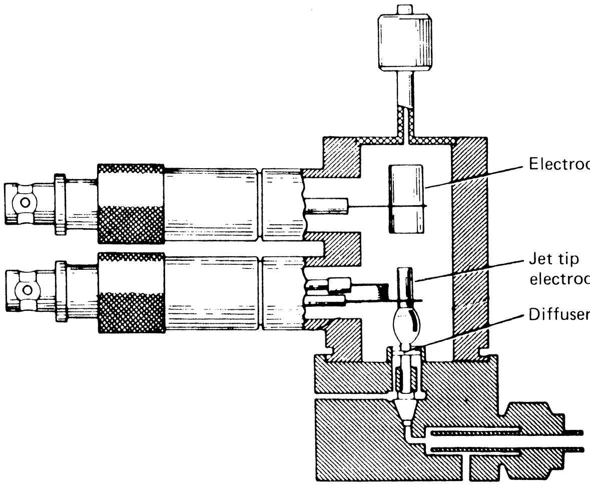


FIGURE 10 The flame ionization detector.

increasing the current in accordance with the solute concentration. With most flame-ionization detectors, this current increase is linear with the solute concentration up to six orders of magnitude.

The flame-ionization detector is a carbon counter; each carbon atom in the solute molecule that is capable of hydrogenation is believed to contribute to the signal (compounds with C—C and C—H bonds), while the presence of nitrogen, oxygen, sulfur, and halogen atoms tends to reduce the response. The detector is most sensitive for hydrocarbons. Practically, no response is obtained for inorganic gases, carbon monoxide, carbon dioxide, and water.

Because of its high sensitivity (the minimum detectable amounts are of the order of 10^{-12} g/s), linearity, and ease of operation, this detector is most popular, in spite of the somewhat incomplete understanding of the physical (ionization) processes involved.

3. The Electron-Capture Detector

This detector is a device based on certain gas-phase ionization phenomena within the ionization chamber. Its schematic diagram is given in Fig. 11. The carrier gas molecules, flowing through the ionization chamber, are bombarded by the radioactive rays from the source of radiation (usually a foil containing ^{63}Ni or ^{3}H) incorporated into the detector body. In a rather complicated process, radicals, positive ions, and low-energy electrons are created. Application of electric potential between the electrodes permits the easily collected electrons to be continuously monitored as the so-called standing current (typically, around 10^{-9} A). This steady current provides a baseline value for the measurement of substances with a strong affinity to such low-energy electrons. When an electron-capturing solute enters the detector, it decreases the population of electrons by an electron attachment process. A decrease of standing current thus occurs during the passage of a solute band, resulting in a negative chromatographic peak.

The decrease of standing current due to the electron-capture process is proportional to the solute concentration in a process reminiscent and formally similar to Beer's law of optical absorption, except that thermal-energy electrons rather than photons are involved:

$$E = E_0 \exp(-Kxc), \tag{12}$$

where E is the number of electrons reaching the anode per second, E_0 is the initial number of electrons, K is the electron-capture coefficient (a function of molecular parameters), x is the detector geometrical constant, and c is the solute concentration.

The electron-capture detector is a selective measurement device since only certain compounds exhibit appreciable affinities toward the low-energy electrons. Among the structures exhibiting strong electron affinities are various halogenated compounds, nitrated aromatics, highly conjugated systems, and metal chelates. The detector is extremely sensitive (amounts between 10^{-12} and 10^{-15} g can be detected) to various pesticides, herbicides, dioxins, freons, and other substances of great environmental concern. To achieve this extremely high sensitivity for normally noncapturing types of molecules (e.g., hormones and drug metabolites), various electron-capturing moieties can be introduced via chemical derivatization (a controlled sample alteration).

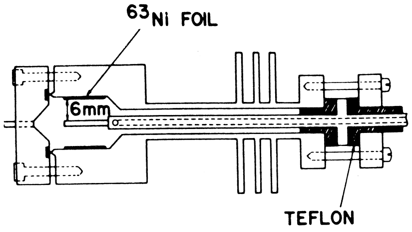


FIGURE 11 Electron capture detector. [From Fenimore, D. C., Loy, P. R., and Zlatkis, A. (1971). *Anal. Chem.* **39,** 1972.]

4. Other Detection Techniques

Several additional detectors were developed for GC. A major aim of such measurement devices is selectivity together with high sensitivity. Selective detectors should be blind to compounds in a mixture that do not possess certain unique structural features (i.e., chromophores or heteroatoms). In practice, some detectors qualify for such selectivity; in other cases, certain substances merely enhance the detector response.

TABLE II Properties of Some Gas Chromatography Selective Detectors

Detector	Selectivity mode	Approximate sensitivity (g)
Electron-capture	Affinity to low-energy electrons	10^{-13}–10^{-14}
Thermionic	Nitrogen	10^{-12}
	Phosphorus	10^{-13}
Flame-photometric	Sulfur	10^{-9}
	Phosphorus	10^{-11}
Electrolytic-conductivity	Halogen compounds	10^{-10}
Ultraviolet	Aromatics	10^{-9}
Photoionization	Partially enhanced response to certain organic molecules as compared with flame ionization (not truly selective)	10^{-11}–10^{-12}

The most common GC selective detectors are listed in Table II together with their analytically important features. These selective detectors have been finding an increasing utilization in the analysis of environmental and biological mixtures. As seen from Table II, sensitivities at the low nanogram level are very common, while some detectors reach levels even below picogram amounts. Parallel uses of a nonselective and a selective detector are quite popular in chemical identification efforts.

B. Ancillary Techniques

While GC is a powerful separation method, it provides only limited information on the chemical nature of the substances it so effectively separates. Consequently, it has to be combined with ancillary techniques. These are certain sample manipulative techniques that are coupled in either a precolumn or a postcolumn arrangement to GC. Their purpose is to enhance qualitative information about the sample, to characterize it chemically, or ideally, to determine unequivocally its structure. Some of these ancillary techniques chemically alter the sample during the process; others measure only its physical parameters, such as optical spectra. On occasion, ancillary tools may represent instruments that are considerably more sophisticated and expensive than the GC instrumentation itself. The three GC ancillary techniques discussed below are among the most powerful and illustrative of this direction.

1. Pyrolysis/GC

This combination is an example of the precolumn arrangement. Pyrolysis/GC combines a controlled thermal degradation of a sample with the subsequent separation of neutral thermal fragments. Most typically, the samples under investigation are large and nonvolatile compounds, such as synthetic or natural polymers. A reproducible pyrolysis/GC process results in the formation of pyrograms that are often highly indicative of some structural details of the original substance; both the presence of certain chromatographic peaks and their areas are judged. As small as submicrogram samples have been successfully analyzed by this combination.

Design of a precolumn pyrolysis unit and the method of thermal degradation are crucial to the acquisition of diagnostically useful pyrograms. Sample size and the pyrolysis temperature must also be carefully controlled. The three most common pyrolysis techniques use (a) filament (ohmical) heating, (b) rapid warp-up of a ferromagnetic conductor in a high-frequency field (Curie-point pyrolysis), and (c) direct thermal degradation in a heated quartz tube. In each case, the sample is deposited from its solution onto a suitable matrix, and the solvent is dried off prior to pyrolysis. Alternatively, small pieces of solids are directly pyrolyzed.

Pyrolysis/GC is used extensively in the analysis of polymers, paints, textile fibers, and even whole microorganisms. Certain materials of forensic interest have been characterized by this approach. A unique pyrolysis/GC system was aboard the Viking 1975 Mission spacecraft to investigate the possible occurrence of organic compounds in the martian soil.

2. GC/Mass Spectrometry

Mass spectrometers are sophisticated instruments that work on the principles of compound ionization and fragmentation (typically through the bombardment by electrons or selected ions), the physical separation of the charged fragments, and their detection. The information obtained by mass spectrometry is a mass spectrum (ion intensity versus mass) that is highly indicative of the sample's original structure, virtually a fingerprint of a molecule. Consequently, the method provides a powerful means to identify various organic compounds but works more effectively with pure substances than with substance mixtures. The combination of GC with mass spectrometry provides an ideal analytical system, in which the complex mixtures are first separated, and the mass spectrometer is permitted to analyze the substances, one at a time.

Commercial instruments that combine the two techniques vary in several respects. The low-resolution instruments provide the designation of nominal molecular weights, while the high-resolution instruments can work up to the precision of a small fraction of such nominal masses. For example, a low-resolution masses. mass

spectrometer "sees" the proton (^{1}H) as the mass 1 and oxygen (^{16}O) as 16; a high-resolution instrument can measure the same species as 1.0078 and 15.9949, respectively. Consequently, the high-resolution instruments are capable of providing measurements of exact elemental composition for various compounds. Different physical principles of mass separation are involved with these instruments. Importantly, both the low- and high-resolution mass spectrometers can be combined with GC. The methods also strongly overlap with respect to the amounts necessary for analysis.

At first, a coupling of GC and mass spectrometry encountered technological difficulties because the gas chromatograph operates at gas pressures above atmospheric pressure, while most mass spectrometers operate at high vacuum. To overcome these difficulties, molecule separators were developed. These devices, working on principles such as molecular effusion, the jet separation effect, and preferential adsorption on a membrane, selectively remove most carrier gas, reduce pressure in the interface, and allow most sample molecules to pass into an evacuated mass spectrometer. The process of coupling GC to mass spectrometry is further aided by modern pumping technology. In fact, modern combination instruments need no molecule separators for capillary columns (typical flow rates around 1 ml/min).

Contemporary GC/mass spectrometry instruments are greatly aided by computers, which can control various instrumental parameters, provide data reduction, and compare acquired mass spectra with the extensive libraries of many thousands of previously recorded spectra.

3. GC/Infrared Spectroscopy

Infrared (IR) spectra of organic compounds are characteristic of various functional groups in the molecules. IR spectral information is somewhat complementary to mass spectral information. Therefore, the combination of GC with IR spectroscopy is, after GC/mass spectrometry, the second most important structural identification tool. Since conventional IR spectroscopy is less sensitive than most GC detectors, the necessary sensitivity enhancement is achieved through the use of Fourier transform techniques. With the advent of refined optical systems and fast computational techniques, the combination of GC with Fourier-transform IR spectrometry is becoming widely used, although its sensitivity is currently less than that of mass spectrometry. Special optical cells were designed for the purposes of this combination.

V. INSTRUMENTATION

The variety of GC analytical applications, columns, and specialized techniques make the modern gas chromatographs quite sophisticated instruments with precise electronic and pneumatic controls.

The carrier gas and the auxiliary gases for detectors are controlled by a set of pneumatic devices (pressure regulators, flow-controllers, and restrictors) to assure (a) reproducibility of the column flow rate, and thus retention times, in multiple analyses; (b) adjustment of the gas linear velocity for optimal column efficiencies; and (c) reproducibility of detector response for reliable quantitative measurements. In addition, filtering devices are inserted in the gas lines to purify all gases mechanically and chemically.

Type and design of the injection port are crucial to performing separations with different types of chromatographic columns. Different physical dimensions of the packed and capillary columns cause substantial differences in the optimum volumetric flow rates. While typical values for conventional capillary columns range around 1 ml/min, various packed columns pass one to two orders of magnitude greater gas flows. The volumes of injected samples must be adjusted accordingly. In a typical sampling procedure with a packed column, liquid samples of up to a few microliters are injected by a miniature syringe, through a rubber septum, into the hot zone of the injection port. Rapid sample evaporation and transfer into the first section of the column are feasible because of a sufficiently high flow rate of the carrier gas.

Considerably smaller samples are necessary for the much narrower capillary columns. Since small fractions of a microliter can be neither reproducibly measured nor easily introduced into the capillary GC system, indirect sampling techniques are employed. In a commonly used sampling method, a sample volume of approximately 1 μl, or slightly less, is injected into a heated T-piece, where an uneven separation of the vaporized sample stream occurs. While the major part of the sample is allowed to escape from the system, a small fraction (typically, less than 1%) enters the first section of a capillary column. Sampling devices based on this principle are called splitting injectors or splitters. They are generally adequate in situations where samples with high concentrations of the analyzed substances are encountered.

Other ways of indirect sampling onto a capillary column involve the injections of (relatively nonvolatile) samples diluted in a sufficiently large (measurable) volume of a volatile solvent (which serves as a sample "vehicle"). With the column inlet kept at a sufficiently low temperature, the nonvolatile sample trace is trapped at the inlet and focused into a narrow zone, while the volatile solvent is allowed to pass through the column and widely separate from the sample. A subsequent increase of temperature permits the sample zone to desorb from its inlet position and enter the usual separation process.

Most sample introduction techniques in GC have now been automated. Process automation permits repeatable

analysis and unattended operation of the instrument. Moreover, reproducibility of the sample injection is improved considerably.

A precise column temperature control is now required for all commercial gas chromatographs. In practice, the GC ovens are designed to have low thermal mass. Resistance spirals situated inside the oven are proportionally heated, while the air circulation throughout the oven is provided by a fan. For adequate analytical work, the column temperature should be reproducible within at least $\pm 0.1°C$.

Retention times in GC are affected by temperature. In accordance with Eqs. (1) and (2), the retention time decreases with increasing temperature because the partition coefficient is decreased. Various solutes, depending on their structures and the chemical nature of a particular phase system, have different dependencies on temperature. Consequently, temperature optimization is necessary for the maximum resolution of the analyzed components. For the mixtures of components with very different values of partition coefficient, column temperature programming is often employed. Commercial gas chromatographs are equipped with convective heating systems that facilitate linear temperature programs. As the column temperature is being gradually raised from a certain initial value to the maximum permissible temperature for an analysis, the sample components with increasingly higher boiling points are eluted from the column. According to the needs of analysis, programming rates are adjustable from as slow as 0.5°C/min up to 30°C/min. Nonlinear and multistep temperature programs are also feasible for special applications.

Modern GC utilizes sensitive detectors. As the measured detector signals (changes in current, voltage, etc.) are quite small, electronic signal amplification is necessary. Since the gas chromatographs are further provided with integrating devices and small computers to calculate exact retention times and peak areas for quantitation, the signals are converted to their digital forms. In addition to displaying a chromatogram on a recorder, modern GC instruments are capable of performing some advanced tasks, such as computing relative retention values, adjusting the detector baseline, and performing certain forms of data reduction.

VI. APPLICATIONS

Gas chromatography is a highly developed analytical method. It has found great use in the routine analysis of various mixtures of organic compounds. Quantitative GC measurements can frequently be carried out with a remarkable degree of reproducibility (analytical error within a few percent). For accurate determinations, it is advisable to use appropriate standard compounds. Some automated analyzers based on GC principles are also used in the process control and continuous analysis of industrial streams. Specialized GC techniques find their place in scientific research. The extremely high sensitivities of some GC detectors are unparalleled.

The GC method is employed for a variety of mixtures, ranging from permanent gases up to molecules that are almost as large as 1000 Da of molecular weight. The variety of chromatographic columns and detectors available to GC continues to expand its applicability to various analytical problems. Several representative examples will now be described to demonstrate the method's versatility, resolving power, selectivity, and sensitivity. These examples have been chosen from the areas of industrial analysis, occupational hygiene, and biochemical research. Other major areas, not covered here, are geochemistry, food and aroma analysis, various agricultural and environmental analyses, atmospheric measurements, and forensic investigations.

The analysis of light gases (permanent gases, gaseous oxides, and C_1–C_5 hydrocarbons) has been traditionally performed in gas–solid chromatographic systems. Various porous adsorbents possess the capability to adsorb and separate these relatively small molecules. An example is shown in Fig. 12, where the carbon molecular sieve column (6 ft × 1/8 in. i.d.) rapidly resolved a mixture consisting of air, methane, carbon dioxide, acetylene, ethylene,

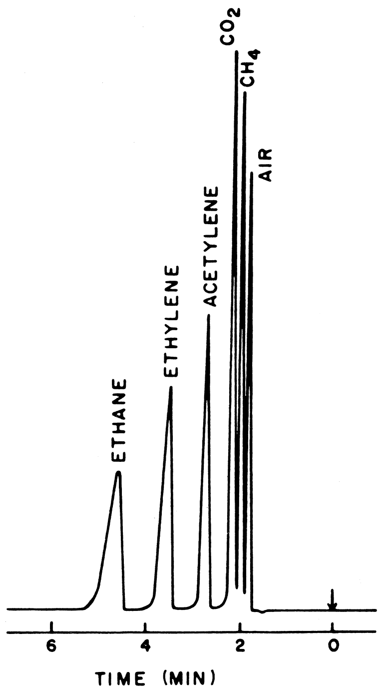


FIGURE 12 Separation of light gases on a carbonaceous adsorbent. [From Zlatkis, A., Kaufman, H. R., and Durbin, D. E. (1970). *In* "Advances in Chromatography 1970" A. Zlatkis, ed. Chromatography Symposium, University of Houston, Texas, p. 120.]

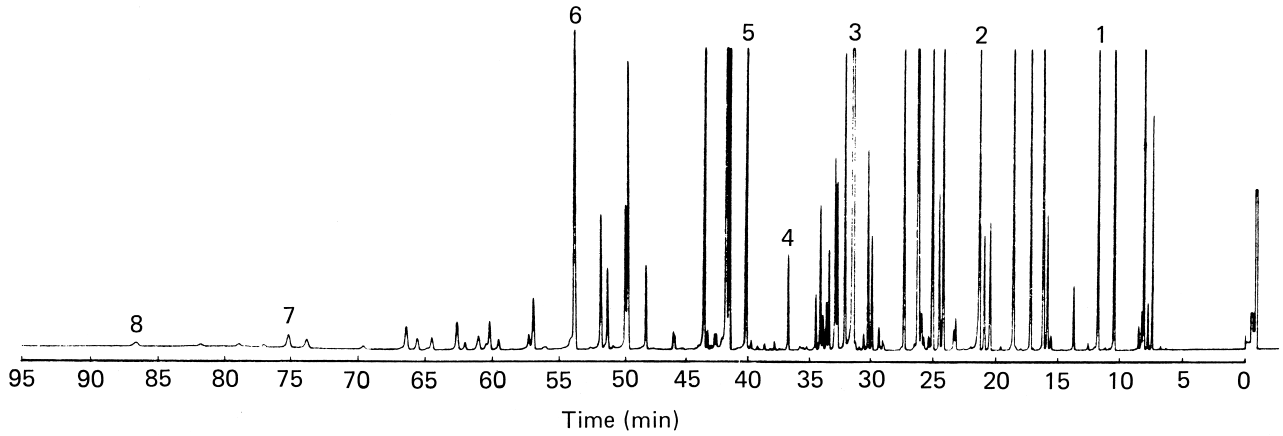


FIGURE 13 Capillary chromatogram of a gasoline sample. [From Adlard, E.R., Bowen, A.W., and Salmon, D. G. (1979). *J. Chromatogr.* **186,** 207.]

and ethane at 150°C (thermal conductivity detection was used). In particular, the separation of acetylene and ethylene is industrially important. While air (a mixture of two major components, oxygen and nitrogen) is eluted here as a single peak, there exist other GC adsorbents that can separate oxygen from nitrogen.

The petrochemical industries have long utilized GC as the analytical method for characterization of various fossil fuels, in monitoring the efficiency of distillation procedures, cracking processes, various chemical conversions, identification of oil spills, and so on. Most samples of petrochemical interest are very complex, so the highly efficient capillary columns are frequently utilized. An example of major-component analysis is shown in Fig. 13, where a full-range gasoline sample has been resolved into a substantial number of components. A 70-m long capillary column was employed, the column temperature was programmed from 0 to 95°C, and the flame-ionization principle was used in the peak detection. A neat gasoline sample was injected (using a sample-splitting technique).

Environmental pollution is among the chief concerns of our industrial society. Highly sensitive analytical methods

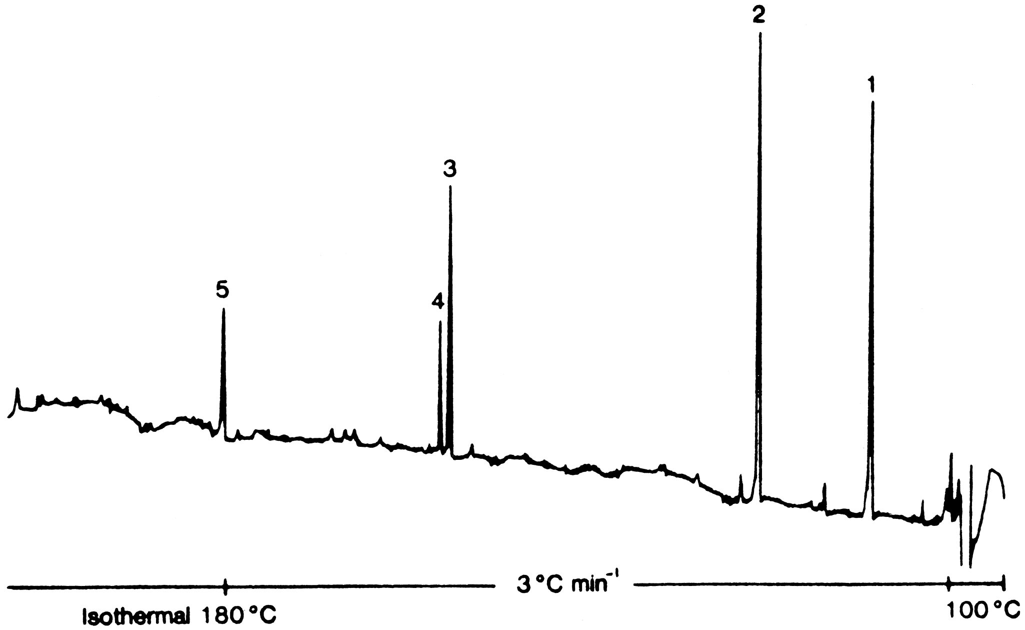
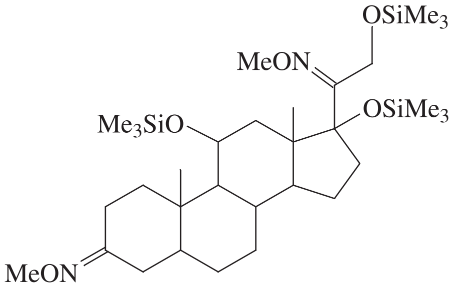


FIGURE 14 Chromatogram of trace aromatic amines (after preconcentration) from the atmosphere of a film-processing laboratory. [From Becher, G. (1981). *J. Chromatogr.* **211,** 103.]

have been developed over the years to identify the sources of water and air pollution, to study biodegradation and transformation of various pollutants in the environment, and to monitor their levels on a continual basis. A great majority of such methods involve GC measurement principles: both packed and capillary columns, selective detectors, the gas chromatography/mass spectrometry combination, and so on. For example, in film-processing laboratories air analysis must periodically be carried out to measure the levels of toxic aromatic amines. With a capillary column (Fig. 14) and the nitrogen-sensitive flame-based detector, five different aromatic amines can be quantitated at the airborne levels of 3 to 13 $\mu g/m^3$.

Prior to the GC analysis, the air sample is first concentrated by passing it through a small adsorbent column. Such a preconcentration step is common if trace organics are to be analyzed in dilute media (air, water, soil, etc.).

Gas chromatography has been applied to analyze numerous biologically important substances such as fatty acids, amino acids and peptides, steroids, carbohydrates, and prostaglandins. Since these compounds are mostly polar and nonvolatile, chemical modifications (sample derivatization) are necessary to block the polar groups and thus enhance volatility of such compounds. To ensure the necessary reliability of GC analyses, such chemical modifications must have highly reproducible yields. For compounds with diverse functional groups, multiple derivatizations (through more than one reaction) are needed. Examples of these are the various steroid hormone metabolites that feature ketonic and hydroxy functional groups in their molecules. Prior to their GC analysis, these compounds are first subjected to treatment with methoxylamine hydrochloride (to form methoximes from ketones) and then to reaction with a trimethylsilyl donor reagent (to form trimethylsilylethers from alcohols). An example of a fully derivatized steroid is a methoxime-trimethylsilyl product of the glucocorticoid hormone, cortisol.



Me = methyl

Other steroids (i.e., compounds structurally related to cortisol) can also be derivatized in a similar manner and subjected to GC analysis. If a high-resolution (capillary) column is employed for their separation, entire profiles of closely related substances can be monitored under different circumstances of health and disease (Fig. 15). While this demonstrated case has been related to an effort to improve our understanding of hormonal alterations in human diabetes, similar analytical GC techniques have been employed to detect abnormalities in adrenal function and reproductive processes.

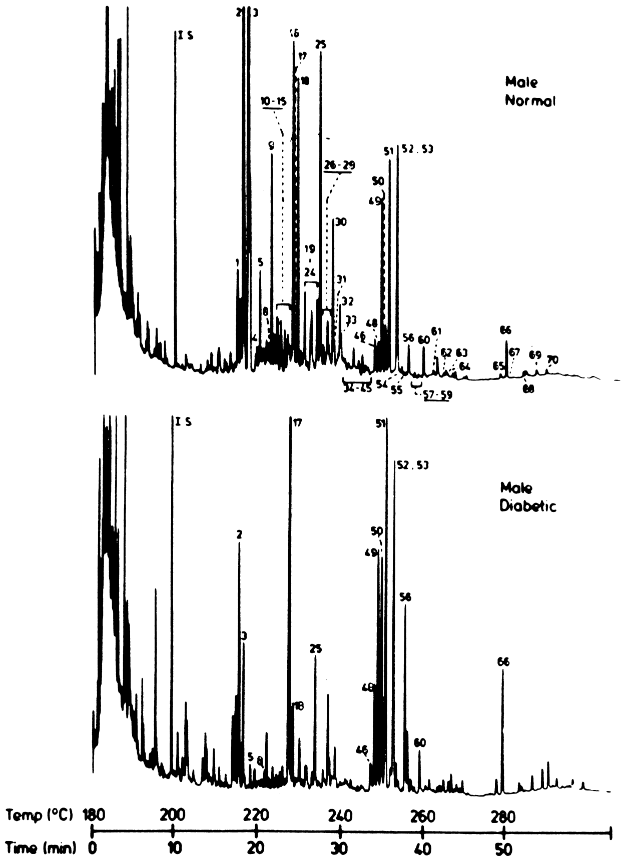


FIGURE 15 Urinary steroid profiles (after sample derivation) from a normal and a diabetic patient, as recorded by capillary chromatography. [From Alasandro, M., Wiesler, D., Rhodes, G., and Novotny, M. (1982). *Clin. Chim. Acta* **126,** 243.]

Gas-chromatographic methods are widely used to analyze amino acids in the hydrolyzates of small protein samples. The method's sensitivity is the major reason for these applications. In addition, GC-based techniques provide opportunities to distinguish and quantitate amino acids (and several other compound types) as different optical isomers. The most popular procedure to separate R and S isomers employs an optically active (chiral) stationary phase. Because of the zwitterionic nature of amino acids, a two-step derivatization is necessary prior to GC. As the first step, the acid (carboxy) function is blocked through esterification. During the second treatment, the amino groups are acylated. Figure 16 demonstrates

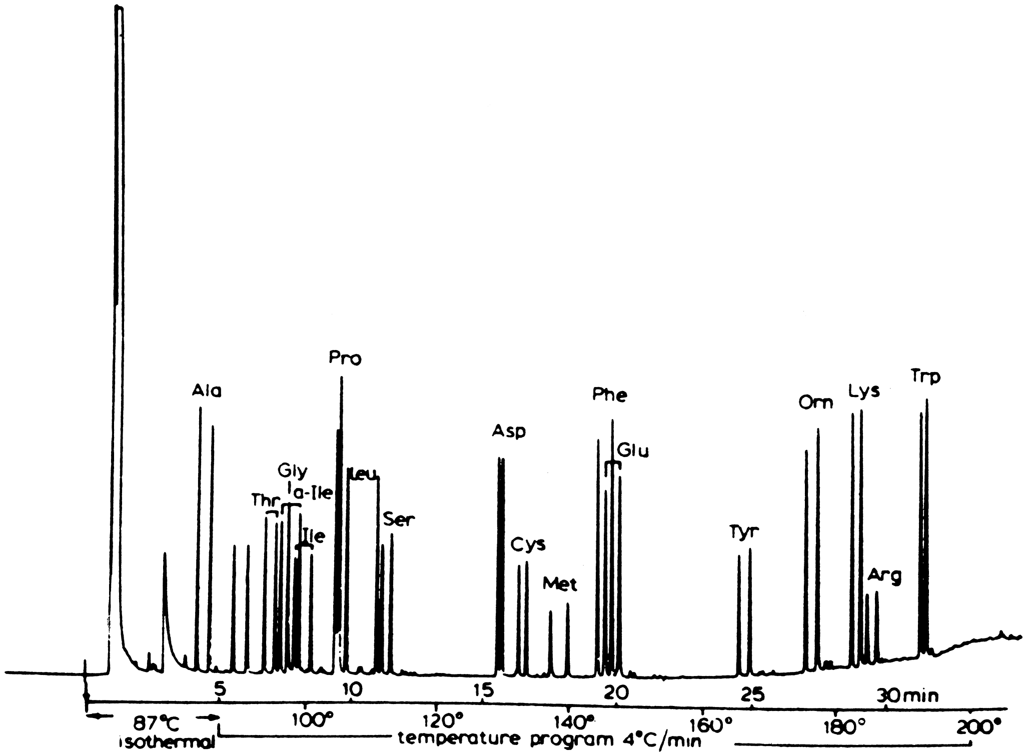


FIGURE 16 Capillary GC separators of a racemic mixture of 19 amino acids on an optically active stationary phase. [From Frank, H., Nicholson, G. J., and Bayer, E. (1978). *J. Chromatogr.* **167,** 187.]

a chromatogram of the 19 naturally occuring amino acids, separated in the form of their volatile derivatives into their respective enantiomers.

SEE ALSO THE FOLLOWING ARTICLES

Electrophoresis • Liquid Chromatography • Organic Chemistry, Compound Detection

BIBLIOGRAPHY

Ettre, L. S., and Zlatkis, A. (eds.). (1984). "The Practice of Gas Chromatography," Wiley (Interscience), New York.

Lee, M. L., Young, F. J., and Bartle, K. D. (1984). "Open Tubular Column Gas Chromatography," Wiley, New York.

Novotny, M., and Wiesler, D. (1984). *In* "New Comprehensive Biochemistry," (Z. Deyl, ed.), Vol. 8, p. 41. Elsevier, Amsterdam.

Poole, C. F., and Schute, S. A. (1984). "Contemporary Practice of Chromatography," Elsevier, Amsterdam.

Infrared Spectroscopy

Norman B. Colthup
American Cyanamid Company

I. Basic Theory
II. Instrumentation
III. Sample Handling Techniques
IV. The Near Infrared Region
V. Quantitative Analysis
VI. Group Frequencies

GLOSSARY

Absorbance Vertical coordinate used for infrared spectra that is equal to the log of the reciprocal of the sample transmittance.

Absorptivity Constant characterizing the capacity of a sample to absorb radiation of a specific wavelength, independent of sample thickness or concentration.

Dipole moment Magnitude of the positive or negative charge constituting a dipole, multiplied by the spacing between the charges.

Fermi resonance Quantum mechanical interaction between close-lying energy states of a fundamental and an overtone or combination that shifts the absorption frequencies and redistributes the intensities.

Fundamental Band in the infrared spectrum that results from a change from the vibrational ground state to the first excited vibrational state of a molecule.

Group frequency Frequency region of the spectrum where absorption is expected when a chemical functional group such as a carbonyl is present in a molecule.

Interferometer Device that splits radiation into two beams and then recombines them, resulting in interferences that depends on the path length difference between the two beams.

Monochromator Device that separates radiation having a single frequency from radiation having many different frequencies.

Normal coordinate Single coordinate defined in such a way that it describes the effective amplitude of a normal mode of vibration in a molecule.

Normal mode Vibration where each Cartesian coordinate of every atom in the molecule oscillates with the same frequency and goes through the equilibrium point at the same time; there is no molecular translation or rotation.

Oscillator Mass or group of masses that vibrate at certain frequencies.

Overtone Band in the infrared spectrum that results from a change from the vibrational ground state to the second or a higher vibrational state of a molecule.

Photon Individual particle or quantum of radiation.

Transition Change from one quantum mechanically defined energy state to another.

THE INFRARED (IR) spectrum results from the interaction of radiation with molecular vibrations and, in gases, with molecular rotations. The spectrum itself is a plot of sample transmission of IR radiation as a function of wavelength or related units. Infrared spectroscopy is the physics that deals with the theory and interpretation of this spectrum and is one of the most popular techniques for identifying molecules. The IR spectrum can be used as a type of "fingerprint" unique to a molecule. In addition, the presence or absence of many chemical functional groups such as phenyls and carbonyls usually can be established from the spectrum. Quantitative analyses of mixtures can be obtained. Infrared spectra can be run for liquids, solids, or gases without special difficulties. Different types of spectrometers can be used, and a wide variety of sample handling techniques are available, many of which are described in this article.

I. BASIC THEORY

A. Electromagnetic Spectrum

Electromagnetic radiation can be characterized by its wavelength λ, its frequency ν, or its wave number $\bar{\nu}$. In the IR region the unit used for wavelength is the micrometer (μm). The frequency unit is cycles per second or hertz (Hz). The wave number unit is cycles per centimeter or reciprocal centimeters (cm^{-1}). The wave number (cm^{-1}) is the number of waves in a continuous wave sequence 1 cm long. The relationship between the units is given in Eq. (1):

$$\bar{\nu}\,(\mathrm{cm}^{-1}) = \frac{1}{\lambda\,(\mathrm{cm})} \qquad \bar{\nu}\,(\mathrm{cm}^{-1}) = \frac{10^4}{\lambda\,(\mu\mathrm{m})}$$
$$\bar{\nu}\,(\mathrm{cm}^{-1}) = \frac{\nu\,(\mathrm{Hz})}{c\,(\mathrm{cm/sec})}. \qquad (1)$$

From this, one can see that wave number (cm^{-1}) is equal to the reciprocal of the wavelength (cm) or is equal to 10^4 times the reciprocal of the wavelength (μm). The wave number (cm^{-1}) in a vacuum is also equal to the frequency (Hz) divided by c, the velocity of light in a vacuum given in centimeters per second. This makes the wave number proportional to the frequency.

The visible region of electromagnetic radiation extends from about 0.38 to 0.78 μm. The IR region extends from the end of the visible region at 0.78 μm to the microwave region with a wavelength of ~1 mm. The IR region is usually divided into three sections. The section used most by chemists is the mid-IR region extending from 2.5 μm, or 4000 cm^{-1}, to ~50 μm, or 200 cm^{-1}. The division at 4000 cm^{-1} is the high wave numberlimit for fundamental vibration absorption in the IR. The low wave number limit is more variable since it is more or less an instrumental limitation. The region between the visible and the mid-IR regions is called the near-IR region. This region of the IR has been used for many applications, especially for quantitative analysis. The region beyond ~50 μm (200 cm^{-1}) is called the far-IR region. This region is used for studying low-frequency vibrations and some molecular rotations.

Electronic transitions give rise to absorption in the ultraviolet and visible regions of the spectrum, and pure rotations of gaseous molecules give rise to absorption in the far-IR and microwave regions of the spectrum. Intramolecular vibrations of molecules give rise to absorption throughout most of the IR region.

B. Infrared Spectra Coordinates

Some examples of IR spectra are given in Figs. 1 through 6. Chemical group vibrations associated with spectral bands are indicated. These are all liquids run in a 0.01-mm-thick NaCl cell. The horizontal coordinates for IR spectra are usually either linear with wavelength (μm) or linear with wave number (cm^{-1}) with generally a factor of 2 scale change at 2000 cm^{-1}. An advantage of the wave number scale is that the wave number of the radiation is proportional to its frequency and to photon energy, and these properties are related to the frequencies and energies of molecular vibrations.

The vertical coordinate in a single-beam spectrum is a measure of the intensity of the radiation of a given wave number that has passed through a sample and reached the detector. Usually, this spectrum is ratioed with another single-beam reference spectrum without a sample to give a ratioed or double-beam spectrum. The transmittance T is the intensity of the sample single-beam spectrum divided by the intensity of the reference single-beam spectrum at the same wave number. The vertical coordinate most commonly seen in an IR spectrum is linear with percent transmittance (%T), which is transmittance T multiplied by 100.

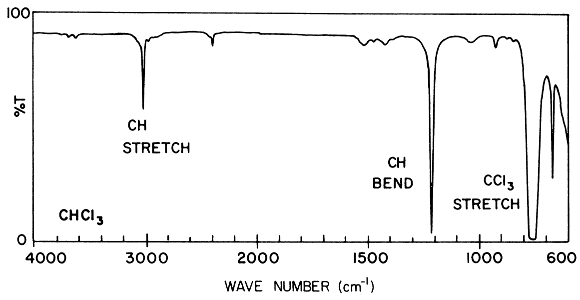


FIGURE 1 Infrared spectrum of chloroform in a 0.01-mm NaCl cell.

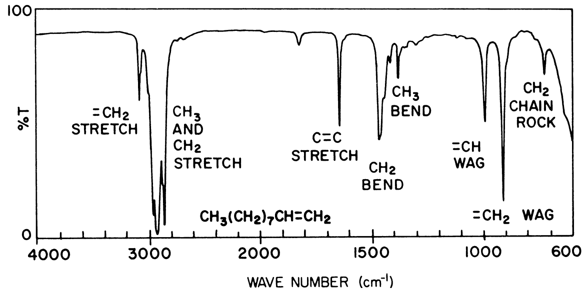


FIGURE 2 Infrared spectrum of 1-decene in a 0.01-mm NaCl cell.

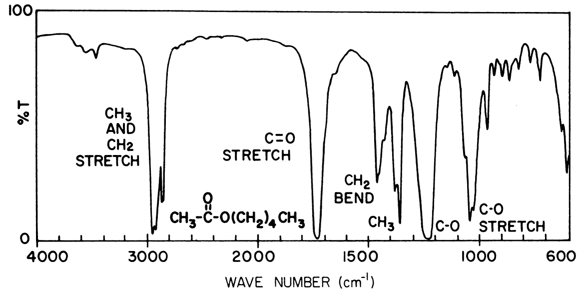


FIGURE 4 Infrared spectrum of *n*-amyl acetate in a 0.01-mm NaCl cell.

Another vertical coordinate scale used increasingly is the linear absorbance scale, where absorbance A is given by the $\log_{10}$ of the transmittance reciprocal:

$$A = \log_{10}(1/T). \tag{2}$$

The advantage of an absorbance scale is that the absorbance is proportional to the product of sample thickness and concentration. This is discussed in section V.A.

C. Diatomic Vibrations

When molecular vibrations are studied, it is useful to consider the nucleus of any atom in the molecule as a mass concentrated at a single point, held in place by chemical bonds that act much like massless coil springs. The diatomic molecule is the simplest molecule type, consisting of two nuclei connected by a chemical bond formed by electrons. The electrons in a molecule move much more rapidily than the nuclei, so the electrons can quickly equilibrate into new electronic structures as the slowly moving nuclei change their spacing. This means that there is a definite potential energy for each nuclear configuration. For the diatomic molecule this is something like the potential energy of an ordinary coil spring. There is a certain equilibrium bond length where the energy is a minimum. If this length is increased or decreased, the potential energy increases and a restoring force is generated, tending to restore equilibrium. In the diatomic molecule, there is only one bond and only one vibration, that which periodically changes the length of the bond. In the harmonic oscillator approximation, the restoring force on each mass is assumed to be linearly proportional to the bond length change from the equilibrium bond length. The proportionality constant is called the force constant k. The vibrational frequency ν of the diatomic molecule can be calculated by the methods of classical mechanics, as in Eq. (3):

$$\nu = \frac{1}{2\pi}\left[k\left(\frac{1}{m_1} + \frac{1}{m_2}\right)\right]^{1/2}. \tag{3}$$

Here ν is the vibrational frequency, m_1 and m_2 the two atomic masses, and k the force constant, namely, the restoring force on either mass divided by the bond length change from equilibrium at any time. Note that the frequency is independent of the vibrational amplitude. The masses move in a manner that keeps the center of mass stationary. If a_1/a_2 is the relative amplitude of the two masses m_1 and m_2 during the vibration, then

$$\frac{a_1}{a_2} = -\frac{m_2}{m_1}. \tag{4}$$

Stretching force constants are usually expressed in millidynes per angstrom (mdyne/Å); 1 mdyne/Å equals 100 N/m, the equivalent SI unit. Masses are usually expressed in unified atomic mass units (carbon = 12). If these

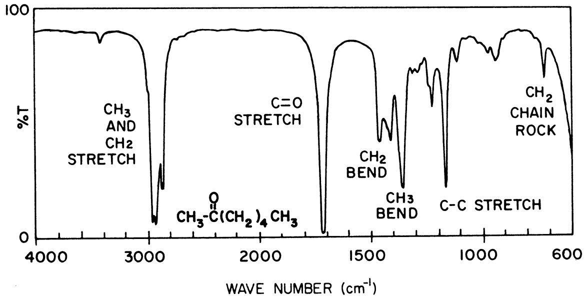


FIGURE 3 Infrared spectrum of 2-heptanone in a 0.01-mm NaCl cell.

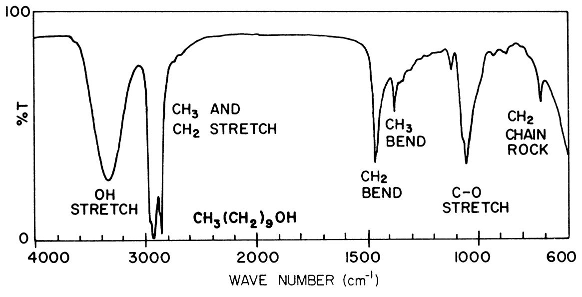


FIGURE 5 Infrared spectrum of *n*-decyl alcohol in a 0.01-mm NaCl cell.

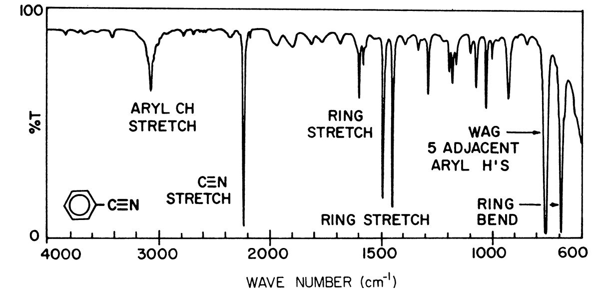


FIGURE 6 Infrared spectrum of benzonitrile in a 0.01-mm NaCl cell.

units are used, the wave number $\bar{\nu}$ of the radiation that has the same frequency as the molecular vibration is given by

$$\bar{\nu} = 1303\left[k\left(\frac{1}{m_1} + \frac{1}{m_2}\right)\right]^{1/2}. \qquad (5)$$

Some examples of diatomic molecules that absorb in the IR are listed in Table I. Equation (5) relates the wave number of the absorption band given to the force constant and atomic masses of the molecule.

D. Polyatomic Vibrations

Polyatomic molecules have more than one vibrational frequency. The number can be calculated from the following. One atom in the molecule can move independently in three directions, the x, y, and z directions in a Cartesian coordinate system. Therefore, in a molecule with n atoms, the n atoms have $3n$ independent ways they can move. The center of mass of the molecule can move in three independent directions, x, y, and z. A nonlinear molecule can rotate in three independent ways about the x, y, and z axes, which pass through the center of mass. A linear molecule has one less degree of rotational freedom since rotation about its own axis does not displace any atoms. These translations of the center of mass and rotations can be performed with a rigid molecule and do not change its shape or size. Substracting these motions, there remain $3n - 6$ degrees of freedom of *internal* motion for nonlinear molecules and $3n - 5$ for linear molecules. These

TABLE I Diatomic Molecules

Molecule	Wave number (cm^{-1})	Force constant (mdyne/Å)
CO	2143	18.6
NO	1876	15.5
HF	3962	8.9
HCl	2886	4.8
HBr	2559	3.8

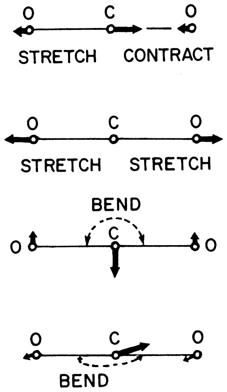


FIGURE 7 Normal modes of vibration of carbon dioxide.

internal degrees of freedom change the size or shape of the molecule without rotating it or translating its center of mass.

It can be shown by the methods of classical mechanics that the $3n - 6$ (or $3n - 5$) internal degrees of freedom of motion correspond to $3n - 6$ (or $3n - 5$) different normal modes of vibration. In a normal mode of vibration the Cartesian displacement coordinates of every atom change periodically, each oscillating with the same frequency and passing through the equilibrium configuration at the same time. The molecule does not translate its center of mass or rotate.

The vibrational form can be described by specifying the relative amplitudes of the Cartesian displacements of each mass (Fig. 7). The vibration can also be described in terms of the relative changes in the *internal coordinates* of the molecule, namely, changes in the bond lengths and bond angles. For example, in Fig. 7 the CO_2 molecule has two bonds; in one vibration, both bonds stretch at the same time (inphase stretch), whereas in another vibration one bond stretches while the other bond contracts (out-of-phase stretch). There are two mutually perpendicular bending vibrations that have the same frequency.

For each normal mode of vibration, a single coordinate can be defined called a *normal coordinate*. When one normal mode of vibration is activated, one normal coordinate periodically changes in value. At the same time that the normal coordinate changes, each Cartesian displacement coordinate changes in a specified proportion (positive or negative) to the change in the normal coordinate, so that the resulting motion is a normal mode of vibration. Normal coordinates are very useful for theoretical studies of molecular vibrations.

E. Infrared Absorption

In a spectrometer, a source of IR radiation sends all IR wavelengths of interest through a sample. The IR radiation

causes some of the molecules to vibrate with increased amplitude, which increases the vibrational energy. The increase in vibrational energy is at the expense of radiation energy, resulting in the absorption of IR radiaition at certain frequencies. The absorption frequency of a fundamental absorption band in the IR is the same as the frequency of the molecular vibration that caused the absorption. This provides a means of observing molecular vibrational frequencies that can provide a molecular "fingerprint" for identifying molecules. The frequencies can also be used to characterize internal features within the molecule that can provide information about the molecular structure.

F. Dipole Moment Change

There must be some means by which the radiation energy can be transferred to the molecule when the molecule absorbs radiation energy. This involves the dipole moment of the molecule. A dipole consists of a positive and a negative charge of equal magnitude separated by a distance. The dipole moment is the magnitude of either charge multiplied by the spacing. Within a molecule, we can picture atoms as particles with small excess negative or positive charges, since chemical forces act to make some atoms have a slight excess or deficiency of electrons. We can picture the negative charge of the dipole as the total excess negative charge of the negative atoms concentrated at the center of the excess negative charge and can picture the positive charge of the dipole as the total excess positive charge on the positive atoms concentrated at the center of the excess positive charge. In CO_2, for example, the center of excess negative charge is between the two electronegative oxygens and the center of excess positive charge is at the relatively electropositive carbon. At equilibrium, these two charge centers coincide with zero spacing, so the dipole moment is zero.

Many molecular vibrations cause the dipole moment to change. For example, in HBr, the bromine is more electronegative than the hydrogen, so the bromine has a slight excess negative charge and the hydrogen a slight excess positive charge. During the vibration, the H–Br spacing changes and also the amount of excess charge on each atom changes, causing the dipole moment to change.

The electromagnetic radiation imposes an electric field on the molecule. This electric field exerts forces on charges, and by definition the forces on positive and negative charges are oppositely directed. The atoms with excess negative charge are pulled in one direction, while the atoms with excess positive charge are pulled in the opposite direction. These forces tend to induce a change in dipole moment. The electric field of the radiation oscillates at the radiation frequency, and this tends to induce an oscillating dipole moment in the molecule. If the radiation-induced dipole moment oscillation has the same frequency as the dipole moment oscillation resulting from a molecular vibration, then the radiation may induce the molecule to vibrate with increased amplitude. If a vibration causes no change in dipole moment, then there is no way the radiation can excite that vibration. In a homonuclear diatomic molecule such as H_2 or Br_2, the two atoms have identical excess charges (namely, zero), and the dipole moment does not change during the vibration as it is always zero. There is no way the electric field of the radiation can induce the two similar atoms to move in opposite directions as required in a vibration. The selection rule for IR absorption requires that in order to be IR-active, a molecular vibration must cause a change in dipole moment. The IR absorption intensity is proportional to the square of the change in dipole moment with respect to the change in the normal coordinate. This reflects the fact that the more the dipole moment changes during a vibration, the greater the probability that the radiation of the proper frequency can excite that vibration.

G. Symmetry and Infrared Activity

If a molecule has some symmetry, a particular vibration of that molecule may be IR-inactive; that is, the vibration will not give rise to any IR absorption. This is because the dipole moment change can be zero as a direct consequence of the symmetry.

One symmetry element is the center of symmetry. If a molecule in the equilibrium configuration has a center of symmetry, one can start at any atom and go in a straight line through the center and an equal distance beyond, where one will find another atom of identical type. An example is carbon dioxide O═C═O (see Fig. 7). Such a molecule at equilibrium would have a dipole moment of zero. The vibrationally distorted molecule where each atom has moved to the end of its displacement vector may have less symmetry than the molecule at equilibrium. One can repeat this procedure for the vibrationally distorted molecule. If it still has a center of symmetry, the vibration is said to be symmetric with respect to the center of symmetry. An example in Fig. 7 is the in-phase stretch of O═C═O, where both CO bond lengths are always equal. Such a vibration is IR-inactive since the dipole moment (which is always zero) does not change. If the vibrationally distorted molecule no longer has a center of symmetry, then the vibration is said to be antisymmetric with respect to the center of symmetry. An example in Fig. 7 is the out-of-phase stretch vibration of O═C═O, where the positive carbon is not a longer midway between the two negative oxygens. The dipole moment changes during this vibration, which is IR active.

Symmetry elements other than the center of symmetry include planes of symmetry and two-fold or higher axes of symmetry. When a plane of symmetry is present, the plane can be thought of as a mirror. When each atom in the molecule is moved to the position of its mirror image, the resulting configuration is indistinguishable from the original. When a twofold axis of symmetry is present, the molecule can be rotated by half a full circle to give a configuration indistinguishable from the original. A full discussion of symmetry and group theory cannot be given here. However, molecules that do not have a center of symmetry may have IR-inactive vibrations as a consequence of these other symmetry elements. For example, the tetrahedral sulfate ion SO_4^{2-} does not have a center of symmetry, but the in-phase stretch of the four SO bonds is IR-inactive. The four negative oxygens move radially at the same time, but the symmetry requires that the center of their excess negative charge does not move relative to the more positive sulfur.

H. Quantum Mechanical Harmonic Oscillator

The simplest classical harmonic oscillator is a single mass m suspended from the ceiling by a spring that obeys Hooke's law. If the mass is pulled down a distance x from its equilibrium point, the spring length minus its length at equilibrium is x. A restoring force on the mass is generated that is proportional to the spring length change. The magnitude of the restoring force equals kx, where k is the force constant. If the mass is held stationary at this point, the potential energy PE is

$$\mathrm{PE} = \tfrac{1}{2}kx^2. \tag{6}$$

This is also the total energy for this condition since the kinetic energy is zero. If the mass is released, it moves toward the equilibrium point and the kinetic energy increases as the potential energy decreases. At equilibrium, the energy is entirely kinetic; the mass overshoots the equilibrium point and continues on until the energy is again entirely potential at maximum amplitude, and the cycle is repeated again. Throughout the vibration the total classical energy is unchanged and is

$$E = \tfrac{1}{2}kx_{\mathrm{max}}^2, \tag{7}$$

where x_{max} is the maximum amplitude. In a classical vibration the maximum amplitude is continuously variable since one is free to pull out the spring to any length before it is released to vibrate. This means that the energy of the classical harmonic oscillator is continuously variable and can have any value.

Oscillators the size of molecules obey the laws of quantum mechanics. The vibrational energy of the quantum mechanical harmonic oscillator is not continuously variable, but has discrete values given from quantum mechanics as

$$E = \left(v + \tfrac{1}{2}\right)h\nu_\circ, \qquad v = 0, 1, 2, \ldots \tag{8}$$

Here E is the vibrational energy, h Planck's constant, $\nu_\circ$ the classical vibrational frequency of the oscillator, and v the quantum number, which can have only integer values. In the classical oscillator, the lowest possible energy is zero when there is no vibration. In the quantum mechanical oscillator, the lowest possible energy is $\frac{1}{2}h\nu_o$, which is not zero, so the molecule can never stop vibrating entirely. This state where $v = 0$ is called the ground vibrational state.

If the vibrational energy is to be increased, the quantum number v must be increased. When the quantum number is increased by 1, the energy change ΔE from the previous equation is

$$\Delta E = h\nu_o. \tag{9}$$

A photon has an energy E given by

$$E = h\nu_{\mathrm{p}}, \tag{10}$$

where ν_{p} is the frequency of the photon. When the photon electric field frequency ν_{p} is equal to the classical dipole moment oscillation frequency ν_{o} for this vibration, the photon will have exactly the right energy (ΔE) needed to increase the vibrational quantum number by 1.

The transition when the quantum number changes by 1 is called an allowed transition in a harmonic oscillator. The most important of these is the transition where the oscillator goes from the $v = 0$ level to the $v = 1$ level. This is called the fundamental transition and is responsible for most of the strong bands in the IR spectrum. The ($v = 0 \rightarrow v = 1$) transition is much more probable than the ($v = 1 \rightarrow v = 2$) transition because at room temperature many more oscillators exist in the low-energy $v = 0$ state than in the $v = 1$ state (or higher states).

In a polyatomic molecule with $3n - 6$ different normal modes of vibration, each normal mode of vibration can be treated separately. In the harmonic oscillator approximation,

$$\mathrm{E} = \left(v_1 + \tfrac{1}{2}\right)h\nu_1 + \left(v_2 + \tfrac{1}{2}\right)h\nu_2 + \cdots \tag{11}$$

where each mode has its own quantum number v and frequency ν. In the harmonic oscillator only one vibration may be excited at one time and the quantum number may change only by 1.

I. Effect of Anharmonicity

In the single-mass harmonic oscillator discussed, the restoring force is a linear function of the mass displacement and the potential energy is a squared function of the mass displacement $\frac{1}{2}kx^2$. Mechanical anharmonicity

results if the restoring force is not a linear function of the mass displacement, in which case the potential energy will have higher-order terms such as cubic and quartic terms. Electrical anharmonicity results if the dipole moment change is not a linear function of the mass displacement. If either mechanical or electrical anharmonicity is present, transitions where the quantum number changes by 2 or more will no longer be forbidden in the IR spectrum. This allows overtones to appear in the spectrum. In a fundamental transition, the quantum number changes by 1 and the photon causing the transition has the same frequency as the classical dipole moment oscillation. In an overtone transition, the quantum number changes by 2 or more. The photon that has the right energy to change the quantum number by 2 has a frequency twice that of the molecular dipole moment oscillation, and in a harmonic-type vibration there will be no dipole moment component changing at this frequency. In an anharmonic vibration, the dipole moment change is complicated by the anharmonicity, and overtones are allowed in the spectrum. The overtone intensity depends on the amount of anharmonicity. Overtones are usually fairly weak.

In a harmonic oscillator, the spacing ΔE between the energy levels for $v = 0, 1, 2 \ldots$ has a constant value hv. If mechanical anharmonicity is present, the spacing is no longer exactly constant, which means that overtone frequencies will not be exactly 2, 3, or more times the frequency of the fundamental. For example, $CHCl_3$ has a CH bending fundamental band at 1216 cm^{-1} and a much weaker CH bending overtone band at 2400 cm^{-1}. A ketone has a carbonyl stretching fundamental band near 1715 cm^{-1} and a much weaker overtone band near 3410 cm^{-1}.

In polyatomic molecules, combination and difference bands are allowed when anharmonicity is present. In a combination-type transition one photon excites two different vibrations at the same time to a new excited state where both vibrational modes have nonzero quantum numbers (say, $v_1 = 1$ and $v_2 = 1$). If both quantum numbers are 1, the combination band will appear in the spectrum near the frequency sum of the two fundamentals. In a difference-type transition, the molecule that is already vibrating in an excited state for one vibration (say, $v_1 = 1$) absorbs a photon of the proper energy and changes to an excited state of a different vibration (say, $v_2 = 1$). The difference band appears at exactly the frequency difference of the two fundamentals in this case. Combination and difference bands, like overtones, are usually fairly weak.

J. Fermi Resonance

In a polyatomic molecule it may happen that an overtone energy level ($v_1 = 2$) has nearly the same energy as a fundamental energy level ($v_2 = 1$) of a completely different vibrational mode. This means that, if no perturbation occurred, an overtone absorption band would have nearly the same frequency as that of the fundamental band of a different vibration. If anharmonicity is present, the higher-order terms in the potential energy may cause perturbations between the fundamental and overtone types involved, generating new mixed energy levels. The vibration types involved should be those that can be coupled by the anharmonic potential function, which requires them to be of the same symmetry type. The perturbation can become significant when the unperturbed energy level difference is small. Combination bands, as well as overtones, can be involved in this interaction, which is called Fermi resonance.

Consider the case where the unperturbed overtone and fundamental nearly coincide. When interaction occurs, two strong bands appear in the spectrum, above and below the expected position of the overtone and the fundamental before interaction. Both bands involve the fundamental and both involve the overtone. The strong intensity of both bands comes from the fact that the fundamental is involved in both bands. The frequency spacing is a function of the perturbation (Table II).

If the expected frequencies of the unperturbed overtone and the interacting fundamental are not identical but are still close to one another in frequency, interaction will not be as strong as before. Two bands of unequal intensity will be seen at again somewhat wider spacing than that for the two unperturbed bands. The stronger band will be nearer the unperturbed fundamental and will involve more of the fundamental vibration. The weaker band will be nearer the unperturbed overtone and will involve more of the overtone vibration. The weaker band will still involve some fundamental vibration, however, which will cause this "overtone" band to be more intense than an unperturbed overtone.

TABLE II Examples of Fermi Resonance

Molecule	Wave number (cm^{-1})	Assignment
NaNCO	620	NCO bend
	(1216, 1305)[a]	(NCO in-phase stretch plus overtone of NCO bend)
	2220	NCO out-of-phase stretch
C_6H_5CHO	1392	Aldehyde CH in-plane bend
	1700	Aldehyde C=O stretch
	(2740, 2825)[a]	(Aldehyde CH stretch plus overtone of CH bend)

[a] The two bands in parentheses have nearly equal intensities, and both involve a fundamental mixed with an overtone of another vibration.

K. Molecular Rotation

Pure rotation of molecules in the gaseous state causes absorption of radiation in the microwave region and to some extent in the far-IR region. In order for a pure rotation to absorb radiation, the rotating molecule must have a permanent dipole moment. Imagine that the dipole moment is oriented perpendicularly to the radiation electric field direction. The field exerts forces in opposite directions on the negative and positive ends of the dipole. This generates a torque, which tends to rotate the dipole moment and therefore tends to increase the rotational frequency of the molecule. If the rotational frequency increases, the rotational energy increases at the expense of the radiation energy. The rotational energy, like the vibrational energy, is not continuously variable but is quantized. Imagine a rotating linear molecule such as HCl in a certain rotational energy state with a quantum number J, which has an integer value (1, 2, 3, ...). The selection rule for pure rotation states that photon energy absorption can increase the quantum number by only 1 to the state $J + 1$. The photon that has the right energy to cause this transition has a frequency intermediate between the classical rotational frequencies for the initial (J) and the final ($J + 1$) states. Unlike the classical vibrational frequency, the classical rotational frequency of the molecule increases during the transition, and the oscillating electric field of the photon with this intermediate frequency can stay nearly synchronized with the rotating dipole moment throughout the transition. An analysis of the rotational fine structure in the spectrum may yield information about the moments of inertia of the molecule.

L. Gas-Phase Band Contours

In the vibrational spectrum the molecule usually changes from the ground vibrational state ($v = 0$) to the first excited vibrational state ($v = 1$). When the sample is in the gaseous state, the molecule may change its rotational state at the same time it changes its vibrational state. The molecule in the ground vibrational state is rotating with a certain angular momentum. When the molecule ends up at the first excited vibrational state, it may be rotating with an increased or decreased angular momentum. As a result of the rotational energy changes, rotational structure is superimposed on the vibrational band, which is referred to as a vibration–rotation band.

If the molecular moments of inertia are sufficiently low and the spectrometer has adequate resolution, rotational fine structure can be resolved in the vibration–rotation band. For larger molecules, the fine structure is usually unresolved, resulting in a broad band. The contour of the band may reveal the direction of the dipole moment change

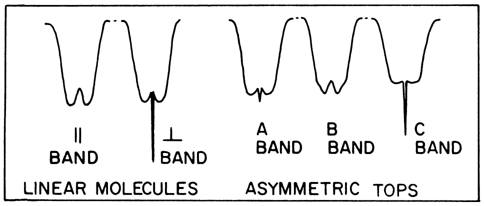


FIGURE 8 Infrared spectra of gas-phase bands for linear molecules and asymmetric top molecules. Unresolved contours are shown for different types of bands. Asymmetric top molecules have different contours for different ratios of the moments of inertia. The parallel band of linear molecules and the B-type band of the asymmetric top have no central peak.

caused by the vibration. In polyatomic linear molecules such as CO_2 and acetylene, IR-active stretching and bending vibrations cause dipole moment changes parallel and perpendicular to the molecular axis, respectively. For parallel vibrations, the gas-phase band contour is a broad doublet (Fig. 8). In the low-frequency and high-frequency wings of the band, the rotational quantum numbers for each of the various energy states have decreased and increased by 1, respectively, during the vibrational transition. Perpendicular bands have the same broad doublet seen in the parallel bands, but a perpendicular band has an additional central peak not seen in the parallel band, where the rotational energy remains unchanged during the vibrational transition.

Tetrahedral or octahedral molecules such as CH_4 and SF_4 are called spherical tops and have three equal moments of inertia for rotation about three mutually perpendicular axes. The gas-phase contour is similar to the perpendicular band of the linear molecule, with two broad wings and a central peak for all the IR-active vibrations.

Molecules with one threefold or higher axis of symmetry such as $CHCl_3$, BF_3, and C_2H_6 are called symmetric tops. Two moments of inertia are equal and differ from the third unique moment of inertia for rotation about the axis of threefold or higher symmetry. When a vibration causes a dipole moment change parallel to the major symmetry axis, the unresolved gas-phase contour is similar to the perpendicular band of the linear molecule, a broad doublet with a central peak. When the dipole moment change is perpendicular to the major symmetry axis, the band structure is more complex and the unresolved contour depends on the relative magnitudes of the moments of inertia.

In molecules with less symmetry, the three moments of inertia are different. These are called asymmetric tops. The axes with minimum and maximum moments of inertia are called the a and c axes, respectively, and the axis with intermediate moment of inertia is called the b axis. Vibrations with dipole moment changes parallel to these axes are called A, B, and C bands. The unresolved

gas-phase contours of A and C bands have broad, more or less symmetric wings and a sharp central peak. The B band has more or less symmetric wings, but is unique in that it has no central peak. The contours of all these bands are dependent on the relative values of the moments of inertia. If the moment of inertia for rotation about the *c* axis is relatively large, as in planar molecules, the central peak of the C band is relatively strong compared with the central peak of the A band, as seen in Fig. 8. If the dipole moment change is not exactly parallel to any of the *a*, *b*, or *c* axes, a mixed contour results.

II. INSTRUMENTATION

A. Infrared Spectrometers

Infrared spectrometers come in a variety of types but have many common features. All have a source that emits all the IR radiation of interest. These are usually various solid materials heated to incandescence by an electric current. The radiation energy distribution as a function of wavelength approaches that of a theoretical black body where the energy reaches a maximum at a wavelength (μm) equal to $2897/T$, where T is the absolute temperature (K). The operational temperature is such that the radiation energy is usually at a maximum near the short-wavelength limit of the spectrum (usually $\sim$2 μm) and decreases as the wavelength gets longer. In the far-IR region, source energy is very low.

All spectrometers must have some kind of detector. These are devices that, in one way or another, change radiation energy into an electrical signal that can be amplified and processed to yield a spectrum. Thermal detectors measure the heating effect of the radiation and respond equally to all wavelengths. Examples include thermocouples, bolometers, and pyroelectric detectors. Detectors that utilize photon energy to free bound electrons in the detector material are called photodetectors. Photodetectors, unlike thermal detectors, do not respond to all wavelengths but have a long-wavelength limit where the photon has insufficient energy to excite the electrons. One example is the photoconductive detector, in which the absorption of photon energy promotes bound electrons to free states. This results in increased electrical conduction.

In between the source and detector, the spectrometer must have some means of analyzing the radiation so that an intensity can be deduced for each wavelength resolution element. Two completely different types of devices are used, namely, monochromators and interferometers. Monochromators with gratings or prisms are used in dispersive instruments, and interferometers are used in Fourier transform instruments.

B. Grating Spectrophotometers

Most spectra seen in the literature are of the ratioed or double-beam type. A double-beam grating instrument is called a spectrophotometer. In this type of instrument the beam from the source is divided into two beams: a sample beam and a reference beam. The sample is placed in the sample beam, and the two beams are alternately passed into the monochromator through the entrance slit, usually at 13 Hz.

In a monochromator, radiation from the entrance slit goes to a paraboloidal mirror, which makes the radiation parallel. The parallel radiation goes to a diffraction grating, which consists of a reflecting surface with straight a parallel grooves very closely spaced. Each of these grooves acts as an independent slitlike source of radiation, diffracting it in different directions. The radiation from the grating is focused onto the exit slit, and only radition leaving the grating at the specific angle goes in a direction that can pass through the exit slit (Fig. 9). When radiation leaves the grating at that angle, parallel beams coming from any two adjacent grooves have traveled different distances and, for one particular wavelength of radiation, will be exactly one wavelength ahead or behind one another. This means that beams of this wavelength leaving at this angle from all the grooves will be *in phase* and show constructive interference when they converge at the exit slit. Other wavelengths will not be in phase and will show destructive interference at this angle. This is called the first order. When parallel beams coming from any two adjacent grooves are two, three, or more wavelengths ahead or behind one another, the parallel beams from all the grooves will also be in phase. These are called the second, third, or higher orders. Unwanted grating orders are removed with filters. This means that, for one grating angle, essentially monochromatic radiation leaves

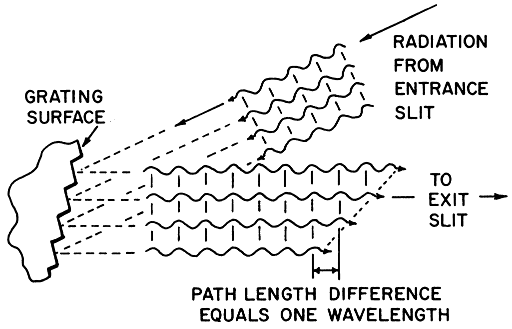


FIGURE 9 The grating surface is shown enlarged with incoming radiation from the entrance slit and reflected radiation going to the exit slit. For adjacent grooves, the beams at these angles have a path length difference. When this equals one wavelength of radiation as shown, all the grooves will emit radiation of that wavelength in phase toward the exit slit.

the monochromator toward the detector. When the grating is rotated to a slightly different angle, the path length difference for beams from adjacent grooves will be slightly different, so radiation with a slightly different wavelength will pass through the monochromator.

When the spectrometer is set at a given wavelength, the sample beam and reference beam alternately pass through the monochromator and activate the detector. If the two beams do not have the same intensity because of sample absorption, an alternating signal is generated and is used to measure the percent transmission of the sample at that wavelength. The grating angle is changed, and the whole spectrum is generated wavelength by wavelength. Usually, several gratings are used for the whole spectral range, and a grating may be used in more than one order. As the wavelength increases, the slit is widened to allow more energy through to compensate for decreased source emission at long wavelengths.

C. Fourier Transform Infrared Spectrometers

In a Fourier transform infrared (FT-IR) spectrometer, there is no monochromator to disperse or separate the radiation by wavelength. Instead, a whole single-beam spectrum is generated all at once. The intensities of all of the wavelength elements are analyzed simultaneously. Since all the radiation frequencies reach the detector at the same time, there is a large signal-to-noise ratio. This is called the multiplex or Fellgett advantage and is one of the principal advantages that an FT-IR spectrometer has over a dispersive instrument. This advantage is particularly noticeable for low-energy conditions or where scale expansion is required to bring out very weak bands. There is also an advantage in that the spectrum can be recorded in less time.

The unique part of an FT-IR spectrometer is the Michelson-type interferometer (Fig. 10). Radiation from the source is made parallel and strikes a beam splitter, typically at 45°. Ideally the beam splitter transmits half the radiation striking it and reflects the other half. One type of beam splitter is a thin layer of germanium coated on an IR-transmitting support. The transmitted and reflected beams leave the beam splitter at right angles, and both strike mirrors, which return the two beams to the beam splitter. The two beams recombine at the beam splitter and show interference. The radiation leaving the beam splitter may go back to the source or may go at right angles, passing through the sample and going on to the detector.

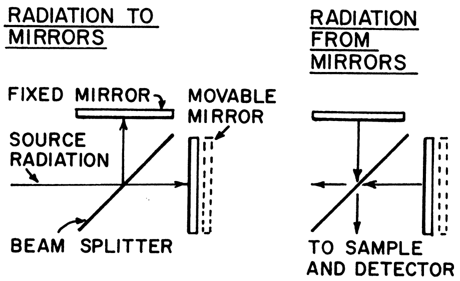


FIGURE 10 Michelson-type interferometer. Left: Source radiation is transmitted and also reflected by a beam splitter to two mirrors. Right: Both mirrors reflect radiation back to the beam splitter, where interference occurs.

One of the two mirrors is movable, so its distance from the beam splitter can be varied. The path length difference for the two beams in the interferometer is called the retardation and is two times the displacement of the movable mirror from the equidistant point. If a monochromatic source such as a laser is used, the radiant energy reaching the detector will vary as a cosine function of the retardation. The detector response will reach a maximum every time the retardation is an integral number of wavelengths of the radiation. At this time the beams from the two mirrors combine at the beam splitter in phase for the beam going to the detector and show constructive interference. If the movable mirror is then moved one-quarter of a wavelength of radiation, the retardation is changed by one-half of a wavelength. The beams from the two mirrors combine at the beam splitter one-half of a wavelenth out of phase for the beam going to the detector and show destructive interference. The detector response as a function of the retardation is called the interferogram. The spectrum can be generated from the interferogram by a Fourier transform. The Fourier transform of a single cosine wave-type interferogram is a single wavelength, in this case that of the laser source.

If a polychromatic source is used, its spectrum can be thought of as a series of closely spaced laserlike emission lines, each with its own wavelength and intensity. Each of these generates a cosine function-type interferogram. The interferogram of a polychromatic source of radiation is a summation of all the cosine functions for each of the laserlike resolution elements. An interferometer does not produce a spectrum but produces an interferogram. A computer must be used to perform the Fourier transform, which generates the spectrum from the interferogram. Once this is done, the computer is available for further processing of the spectrum.

Just one scan of the movable mirror produces a whole single-beam spectrum. However, a spectrum produced from one scan has a relatively high noise level. Usually, a number of scans are taken and signal-averaged by the computer. The noise is reduced by the square root of the number of scans. A single-beam spectrum with the sample in place is stored in the computer memory. A reference single-beam spectrum is taken without the sample and is

also stored in the memory. These two single-beam spectra are ratioed by the computer to give a percent transmittance spectrum.

The computer can be used to modify the spectrum further. For example, the vertical or horizontal scale can be expanded, the background can be straightened, or a linear absorbance scale can be generated. A useful procedure is spectral subtraction whereby, for example, a solvent spectrum can be subtracted from a solution spectrum to yield the pure solute spectrum.

III. SAMPLE HANDLING TECHNIQUES

A. Infrared-Transmitting Materials

One of the features of IR spectroscopy is that solids, liquids, and gases can be run without special difficulties. Usually, some sort of IR-transmitting material is needed to support or enclose the sample. Materials such as glass and quartz are useful as windows in the near-IR but even thin windows do not transmit much below 3000 cm^{-1}. The low wave number transmission limits of IR-transmitting materials are not sharply defined but depend on the window thickness. Four commonly used materials and their approximate low wave number limits are NaCl, 600 cm^{-1}; KBr, 350 cm^{-1}; CsBr, 250 cm^{-1}; and CsI, 200 cm^{-1}. These materials are all water-soluble. Water-insoluble materials and their low wave number limits include CaF_2, 1200 cm^{-1}; BaF_2, 850 cm^{-1}; Irtran-2, 700 cm^{-1}; AgCl, 350 cm^{-1}; and KRS-5 250 cm^{-1}. Irtran-2 is made of zinc sulfide an is often used for water solutions or for making films from water solution. Silver chloride is useful but is soft, deforms easily, and darkens with exposure to light. KRS-5 is thallium bromide iodide and is often used in the internal reflection technique to be discussed later. In the far-IR, high-density polyethylene transmits to as low as 30 cm^{-1} but cannot be used above 600 cm^{-1} because of its absorption.

B. Salt Polishing

Sodium chloride can be easily polished between use. The crystal is sanded flat with a finegrade sandpaper if it is freshly cleaved, scratched, or damaged by water. Two polishing laps are prepared. There are different types, but the wet lap can be simply two thicknesses of fine nylon cloth stretched over a flat surface. The wet lap is wetted with water and sprinkled with a little fine polishing powder such as aluminum oxide or cerium oxide. This is rubbed smooth and all excess water is wiped off. The flat salt plate is rubbed about 20 strokes on the wet lap and then, without delay, is buffed about 7 strokes on the dry lap, which can be simply a layer of diaper cloth held flat. Cesium bromide can be polished the same way, but with alcohol substituted for water. The best polish comes when the lap is nearly dry.

C. Liquid Samples

The easiest samples to run on IR instrumentation are those in the liquid state. Slightly viscous samples can be simply squeezed between two polished IR-transmitting plates and run as a thin film. A typical film thickness is ~0.01 mm. If the liquid is not viscous, usually a spacer is added between the plates to keep the plates apart at the appropriate spacing. Spacer material can be metal foil or an insoluble polymeric film. Two strips roughly 10 by 2 mm can be used, for example, one on each side of the area the IR beam will pass through. These are called temporary cells and are disassembled and cleaned after each use. The thickness cannot be accurately reproduced.

Fixed cells are not disassembled after use but instead are filled, emptied, and cleaned with solvent through ports on the cell assembly. The liquid enters the leak-proof sample area between the plates through holes in the cell window. These are used for volatile liquids or when the thickness needs to be accurately known or held constant as in quantitative analysis. Many commercially available designs are used, and cells come in thickness from 0.01 to 4 mm.

If the cell windows are sufficiently flat, the cell thickness can be measured by running the IR spectrum of the empty cell and observing interference fringes in the form of percent transmission undulations. Wave number $\bar{\nu}_1$ (cm^{-1}) is read at one transmission maximum, and wave number $\bar{\nu}_2$ is read at another transmission maximum that is 1, 2, 3, or more generally n maximum away from the first. The cell thickness t is

$$t\,(\mathrm{mm}) = \frac{5n}{\bar{\nu}_1 - \bar{\nu}_2}. \qquad (12)$$

Interference results because part of the beam is twice reflected inside the cell and is retarded by twice the cell thickness relative to the transmitted beam with which it interferes.

D. Gas Samples

Gas cells used for IR spectroscopy come in a variety of types. The simplest is a basic cylinder 10 cm long with IR-transmitting windows on each end. These may be cemented on or clamped in place, with vacuum-tight gaskets providing the seal. Entrance and exit tubes are provided and fitted with stopcocks. The cell is filled and emptied with a gas handling system.

The sampling chamber of most IR spectrometers is not large enough to accommodate longer cell lengths directly.

However, cells of much longer path length are available that use mirrors to deflect the IR beam and to reflect it back and forth many times in the cell chamber before it leaves the cell and reenters the spectrometer. These long-path-length cells are used for detecting very small quantities of gas in pollution studies, for example.

A technique often used with FT-IR instruments is gas chromatography, or GC–FT-IR. Here the effluent from a gas chromatography column is fed through a heated light pipe with IR-transmitting windows on the ends. Source radiation passes through the cell into the FT-IR spectrometer. The gas chromatography column separates the gas-phase components and ideally sends them one by one through the light pipe, where the high speed of the FT-IR instrument is utilized to get the spectrum of each component "on the fly," so to speak.

E. Solution Spectra

The techniques for running solids in IR are quite varied. In the first case a solid can be dissolved in a suitable solvent and run as a liquid. Unfortunately, no solvent is free of absorption in the IR region and, usually, the better the solvent, the greater its absorption. This means that more than one solvent must be used to get the whole IR solution spectrum in all regions. A commonly used pair of solvents are CCl_4 above 1330 cm^{-1} and CS_2 below 1330 cm^{-1}. These can be used in cells 0.1–1 mm thick, for example, with solute concentrations in the range 10–1%. In double-beam grating spectrophotometers a cell of matching thickness containing solvent only can be put into the reference beam to compensate for the solvent bands. In FT-IR instruments, a reference solvent spectrum can be subtracted from the solution spectrum to remove solvent bands. Another common solvent for solution spectra is $CHCl_3$ often used in 0.1-mm cells with 5 to 10% solute. $CHCl_3$ has strong bands at 1216 and 757 cm^{-1}, where solute information is often lost or inadequately presented. Even water has been used as a solvent for some applications. Here the cell thickness must be kept small, as water is a very strong IR absorber. The internal reflection technique described in Sec. III.I has been successfully used for water solutions.

F. Films

Solid-state films of suitable thickness can be prepared from melts or solution. Such films are most suitable for amorphous materials, especially polymers. Crystalline films may scatter light and show nonreproducible orientation effects from special orientations of the crystal on the IR window surface. A sample can be heated between two salt plates until molten and allowed to solidify. Solutions can be put onto a plate and the solvent evaporated to form a film. This is a good technique for running water-soluble polymers, for example. Sometimes a film can be prepared on a substrate and stripped off and run as an unsupported film. If a film is too uniform in thickness, interference fringes similar to those from an empty cell may be seen, as discussed earlier. If a film is too irregular in thickness, a spectrum with a false percent transmittance will result from the fact that different parts of the beam go through sample areas with different thicknesses.

G. Mulls

One of the best techniques for running crystal-line solids is the use of a mineral oil or Nujol mull. Here a few milligrams of sample are finely ground with a small amount of mineral oil to make a thick paste like cold cream, for example. The paste can be prepared with a mortar and pestle and spread between two IR-transmitting windows. A well-ground sample has a brownish color like smoke when one looks through it. Most beginners do not grind the sample well enough and use too much oil. Mineral oil has only a few bands in narrow regions. The CH stretch region between 3000 and 2800 cm^{-1} and the CH bend region at about 1460 and 1375 cm^{-1} are obscured, however. If information is needed in these regions a second mull must be prepared using a halogenated oil such as Halocarbon or Fluorolube, which contain CF_2 and CFCl groups but no CH. These have no bands from 4000 to 1300 cm^{-1} but have strong bands below 1300 cm^{-1}. Some people use the halogenated oil spectrum above 1300 cm^{-1} and the mineral oil spectrum below. In this case care must be taken to ensure that the sample thickness is the same in both preparations.

H. Potassium Bromide Disks

A very popular technique for running solids is the KBr disk technique. Here a few milligrams of sample are very finely ground and then mixed with 50 to 100 parts of dry KBr powder. The mixture is placed in a special device and compressed into a disk at high pressure. If all goes well, a transparent disk results, which is put into the spectrometer and run. Commercial KBr disk makers are available in many forms. Some are activated with wrenches or levers, while others are used with a hydraulic press. Some can be evacuated, which gives the disk transparency a longer lifetime, but this is not necessary if the disk is used promptly.

Advantages of the disk over the mull include the fact that KBr, unlike mineral oil, has no bands above 400 cm^{-1}. Also, many polymers are more easily ground in KBr. Microsamples are easier to prepare with the KBr disk. The KBr disk has disadvantages compared with the mull, however. The biggest problem is that KBr is hygroscopic,

and bands from absorbed H_2O appear in the spectrum in variable amounts that depend on the technique. One never knows whether the water is in the sample or the KBr preparation. Also, spectra of KBr disks are sometimes less reproducible because of changes in sample polymorphism, which result from the preparation.

I. Internal Reflectance

Internal reflectance results when a beam of radiation inside a material of relatively high index of refraction is reflected from the surface interface between this and a material of lower index of refraction. The angle of incidence α is the angle between the beam and a line perpendicular to the surface interface. If the angle of incidence is small, much of the radiation is transmitted through the surface interface and a little is internally reflected. As the angle of incidence gets larger, a certain critical angle α_c, is exceeded, after which all the radiation is internally reflected from the interface and none is transmitted. The sine of the critical angle, $\sin\alpha_c$ is equal to the ratio n_2/n_1, where n_1 is the higher index of refraction and n_2 is the lower index of refraction on the two sides of the interface.

When the angle of incidence is larger than the critical angle, then the beam in the material with the higher index of refraction penetrates a little into the material with the lower index of refraction in the form of an exponentially decaying wave. It is then reflected back out. When the amplitude of the wave passing through the interface has decayed by a factor of (1/e) or about 37%, the reflective penetration (d), into the material with the lower index of refraction is given by

$$d(1/e) = \frac{\lambda}{2\pi n_1\left[\sin^2\alpha_1 - (n_2/n_1)^2\right]^{1/2}}, \qquad (13)$$

where e is the natural log base, λ is the wavelength of the radiation, and α_1 is the angle of incidence in the material with the higher index of refraction. If the material with the lower index of refraction should absorb part of the radiation penetrating into it, then the internally reflected beam leaving the interface will be attenuated by this absorption. Hence we have the name *attenuated total reflectance* (ATR) for this effect.

As used in infrared spectroscopy, one type of internal reflectance plate (Fig. 11) is made of a high index of refraction material, such as thallium bromide–iodide. The plate is usually a few millimeters thick, and the ends are beveled to allow radiation entry into one end at an angle inside the plate. The beam is multiply internally reflected and zigzags between the surfaces until it leaves at the other end. A sample with a lower index of refraction than the plate is pressed into intimate contact with the plate on one or both sides. The zigzagging beam penetrates a few micrometers into the sample on the plate surface at each reflection. The sample thickness is immaterial as long as it exceeds a few micrometers. The sample contact area should go all the way across the plate so none of the beam can bypass the sample. The lengthwise coverage only affects the attenuation intensity.

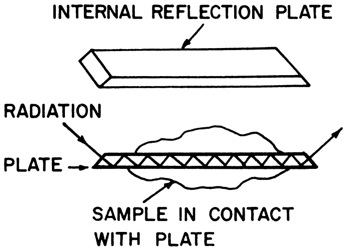


FIGURE 11 Plate used for internal reflection spectroscopy. The lower drawing shows the sample in contact with the plate and radiation being multiply internally reflected within the plate.

When the internally reflected beam is introduced into a spectrometer, the resulting spectrum is similar to a transmission spectrum. There is one major difference. Since the radiation penetration is wavelength dependent in the penetration equation, longer wavelengths penetrate more. The internal reflectance spectrum resembles a transmission spectrum where the sample thickness gets larger in direct proportion to the radiation wavelength.

There is another effect on the penetration, and that is that the index of refraction of the sample (n_2) is not constant. It changes in the region of an absorption band, becoming smaller than average on the high-wave number side of the band center and larger on the low-wave number side. From the penetration equation, the radiation penetration, and therefore the band intensity, will be decreased on the high-wave number side of the band center and increased on the low-wave number side. This distorts the band shape. To avoid this, the denominator in the penetration formula should not get too small. This means that the angle of incidence should not be too small and the index of refraction of the crystal n_1 should be relatively high. In Fig. 11, the angle of incidence can be kept large enough to reduce the band distortion, but this also reduces the band intensity. This is compensated for by using multiple internal reflections.

There have been many variations in the design of internal reflection accessories. In one design, the plate in Fig. 11 is mounted horizontally at the bottom of a shell container so that a liquid can be simply spread over the top of the plate and run without further changes.

In another design, the internal reflectance crystal has a hemispherical shape. In this arrangement, the radiation enters into the curved surface, and is internally reflected off the flat surface, and then exits through the curved surface. In one version of this, the sample is in optical contact

with only a small raised area of the flat surface that allows spectra to be taken of quite small areas. Since there is only one internal reflection, the crystal used, such as silicon, has a very high index of refraction.

There are many applications for internal reflectance spectroscopy, and only a few will be mentioned here. Internal reflectance spectroscopy can be used to obtain the spectra of rubbery materials that are hard to grind. The rubbery material is simply pressed against the internal reflectance plate, and it is ready to run. Carbon-filled rubber or other polymers may be run using a high index of refraction germanium as the internal reflectance element. Internal reflectance is used to obtain selectively the top few micrometers of a sample surface where the composition may be different than that further down. It is also good for water solutions because the controlled penetration keeps the effective sample thickness small.

J. Diffuse Reflectance

Diffuse reflectance is a technique usually used with FT-IR instruments. A powdered sample is placed in a small container, where source radiation strikes it and is diffusely reflected in various directions. This radiation is collected and measured by the spectrometer. Usually in the mid-IR region the finely powdered sample is diluted to 5 to 10% with finely powdered KBr or KCl. The spectrum is ratioed against a reference spectrum of pure powdered KBr or KCl. The ratioed spectrum is processed by a computer using a function $f(R_\infty)$ derived by Kubelka and Munk, which changes the reflectance spectrum into one resembling a linear absorbance spectrum:

$$f(R_\infty) = \frac{(1 - R_\infty)^2}{2R_\infty} = \frac{k}{s} \qquad R_\infty = \frac{R_\infty(\text{sample})}{R_\infty(\text{reference})}. \tag{14}$$

Here (R_∞) is the reflectance of a thick scattering layer, k the molar extinction coefficient, and s a scattering coefficient, which is a function of particle size. The spectrum is quite sensitive to particle size, which affects the radiation scattering. Spectral distortion (compared with a transmission spectrum) may occur if the particle size is not uniformly fine. Black, strongly scattering materials such as coal can be run by this technique.

K. Infrared Microspectroscopy

In this method of sampling, a special type of microscope is used to select very small sample areas for examination by the infrared spectrometer. As in the macroscopic case, samples can be prepared with thicknesses on the order of 0.01 mm. Since infrared radiation must pass through the microscope, all the optics are front surface mirrors. The sampling region can be viewed visually through the microscope, and selected areas can be isolated by masking off the unwanted parts of the field. Variable aperture masks are located in remote image planes of the sample area, located above and below the sample to reduce diffraction effects. Then, the optical path is changed so that the source radiation goes through the unmasked areas and the infrared spectrum of the sample is recorded. This can be divided by the spectrum of a similarly masked blank for example, to give a percent transmission spectrum of the sample, or from this, an absorbance-type spectrum of the sample.

The types of applications are basically similar to those handled by transmission spectroscopy, but with a significant difference. The microscope can yield good spectra on much smaller sample areas. This means that many new types of problems can be handled that were difficult or impossible to solve with standard instruments. Sample areas that are heterogeneous are now easily measured in the chemical industry, forensic work, and biological studies.

In the chemical industry for example, some polymer products may show some very small impurity areas whose chemical composition may be characterized with little difficulty by the spectra. Polymer films weathered by outdoor exposure may show chemical modification as a function of the layer depth below the surface. This may be characterized by the infrared spectra of the various layers.

In forensic work, heterogeneous fields can be examined selectively for hair strands, polymer fibers, or lint particles, for example, and their chemical composition can be characterized by their infrared spectra. Cross-sections of paint chips can be examined and the various layers can be characterized, which may be distinctive.

In the biological field, heterogeneous areas of various biological specimens can be examined, and the infrared spectrum of different microareas of the field can be taken. The spectrum of a single red blood corpuscle can be easily measured. One difference that can be observed in these samples is the ratio of protein to fat. Proteins have bands near 1650 and 1550 cm^{-1} for the O=C-NH group, and fats have an ester C=O band near 1740 cm^{-1} and an alkane chain doublet near 2925 and 2855 cm^{-1}, with an internal unsaturation band near 3015 cm^{-1} for the =CH group. Some studies may reveal chemical differences between adjacent tissues of various types. The spectral characteristics of two-dimensional areas may be mapped by running a series of spectra of a grid pattern. An array consisting of a number of detectors can be used, each one of which generates the infrared spectrum of its own small area of the grid pattern.

IV. THE NEAR INFRARED REGION

A. Qualitative and Quantitative Analyses

The region of the near infrared spectrum extends from about 700 nm (14,000 cm^{-1}) near the end of the visible region, to 2500 nm (4,000 cm^{-1}), which is the beginning of the infrared fundamental region. The HF molecule, for example, has its fundamental absorption band at 3962 cm^{-1}. The near infrared spectrum of water is shown in Fig. 12.

The near infrared region is populated with overtone and combination bands of lower frequency fundamentals below 4,000 cm^{-1}. These bands are considerably weaker in intensity than fundamental bands. The intensity weakness is an advantage for quantitative analyses, particularly in production processing, because the larger cell thicknesses used can be better controlled, yielding more reproduciable intensity measurements. The cells can be made with water-insoluable quartz or saphire windows for example, which if they are thin enough, can transmit down to about 3000 cm^{-1}. On the negative side, the qualitative information is considerably less extensive than in the fundamental region, as most of the near infrared bands involve summation bands (overtones and combination bands) of CH, OH, and NH stretching vibrations. Binary summation bands are usually the most intense, with tertiary and higher summation bands being progressively weaker. Even so, there is still some useful information to be had about functional groups. Characteristic group frequency charts and tables of these group summation bands can be found in the literature. A very brief selection of these are shown in Table III to serve as some examples.

TABLE III Selected Near Infrared Characteristic Bands

Group	nm	cm^{-1}	Assignment[a]
—CH_3	1680–1760	5950–5680	2 x CH_3 str.
—CH_3	2240–2360	4465–4235	CH_3 str. + bend
>C=CH_2	1625–1675	6155–5970	2 x =CH_2 asym. str.
cis–CH=CH—	2120–2170	4715–4610	=CH str. + C=C str.
—C≡C—H	1520–1565	6580–6390	2 x ≡C–H str.
R—OH, H_2O[b]	1385–1460	7220–6850	2 x OH str.
H_2O[b]	1880–1950	5290–5130	H_2O str. + bend
R_2NH	1520–1560	6580–6410	2 x NH str.
R—SH	1960–2020	5100–4950	2 x SH str.
>C=O	1920–1970	5810–5075	3 x C=O str.

[a] str, stretch.
[b] Varies with changes in hydrogen bonding.

B. Derivative Spectra

In the near infrared region, the bands are not only weaker, they can be broader and more overlapping. For example, the near infrared spectrum of oleic acid

$$CH_3(CH_2)_7CH{=}CH(CH_2)_7COOH$$

can be seen in Fig. 13. The overtones of the CH_2 group out-of-phase and in-phase stretch bands show clearly at 1723 and 1758 nm. However, the expected weaker bands for the CH_3 out-of-phase stretch overtone and the =CH stretch overtone, which are expected below 1723 nm, are so broad and weak that they are hard to detect.

For this reason, second derivatives of the spectra are often used, as seen in Fig. 14. Now the strong CH_2 overtone bands at 1726 and 1761 nm are much narrower and better

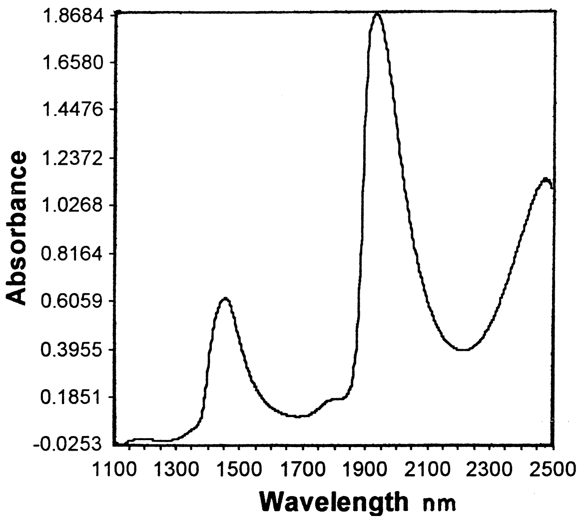


FIGURE 12 The near infrared spectrum of water run in a 0.1-mm cell with bands at 1450 nm (2 × H_2O stretch) and 1950 nm (H_2O stretch + bend).

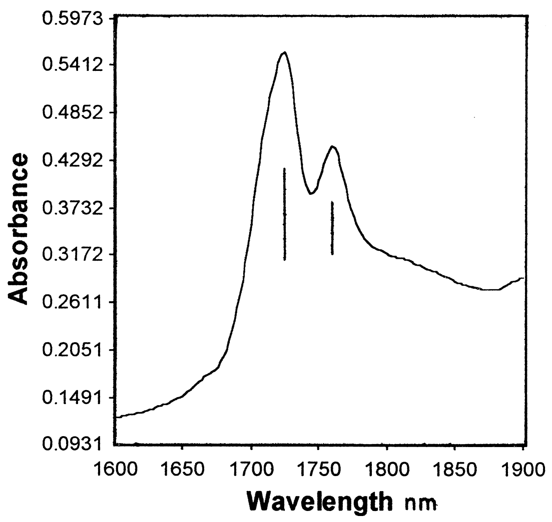


FIGURE 13 The near infrared spectrum of oleic acid run in a 1-mm cell with bands at 1723 nm (2 x CH_2 out-of-phase stretch) and 1758 nm (2 x CH_2 in-phase stretch).

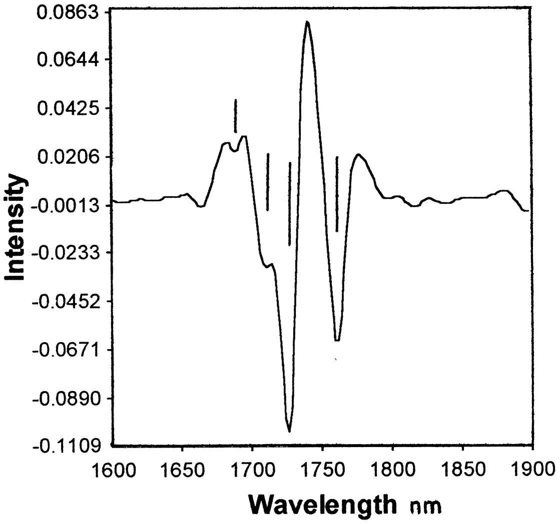


FIGURE 14 The second derivative of the near infrared of oleic acid (Fig. 13). Strong bands are seen at 1726 and 1761 nm (2x CH_2 stretch) as seen in Fig. 13 but with bands pointing down. Bands not seen easily in Fig. 13 are seen here at 1711 nm (2x CH_3 out-of-phase stretch) for the terminal CH_3 and 1688 nm (2x =CH stretch) for the internal unsaturation. Negative side lobes on the sides of the strong bands are seen pointing up.

resolved, and the weak CH_3 and =CH overtone bands are clearly seen at 1711 and 1688 cm^{-1}, respectively. The second derivative of a typical band has a strong central peak, but it also has a weaker negative side lobe on both sides of the main band that complicates the presentation. However, the advantages are that it gives narrower bands that improve the resolution, it brings out weaker bands, and it also straightens out the background.

V. QUANTITATIVE ANALYSIS

A. Beer's Law

The basic law for spectroscopic quantitative analysis is Beer's law. This shows how sample concentration is related to a measure of radiation intensity in a spectrometer. Consider a sample in solution, held in a cell of uniform thickness that transmits the monochromatic radiation of interest. Let the intensity of the radiation entering the sample be I_0 and the intensity of the radiation that has passed through the sample be I. Then the transmittance T is given by

$$T = I/I_0. \tag{15}$$

The percent transmittance ($\%T$) is given by multiplying the transmittance T by 100. Beer's law is commonly expressed as

$$\log_{10}(I_0/I) = abc. \tag{16}$$

Here (I_0/I) can be recognized as the reciprocal of the transmittance $(1/T)$. The quantity b is the cell thickness and c the sample concentration in the solution. The quantity a is called the absorptivity, which is a constant characterizing the capacity of the sample to absorb radiation. Its value varies with the wavelength or frequency of the radiation being measured and with the units used for cell thickness and sample concentration.

Beer's law is more simply expressed as

$$A = abc, \tag{17}$$

where A is called the absorbance, defined as

$$A = \log_{10}(I_0/I) \quad \text{or} \quad A = \log_{10}(1/T). \tag{18}$$

If the cell thickness and the radiation wavelength are held constant, Beer's law states that the concentration is linearly proportional to the absorbance.

B. Beer's Law Deviations

In IR spectroscopy, the concentration range being measured may be large, sometimes ranging from zero to 100%. Under such circumstances, deviations from Beer's law may be observed. If the cell thickness and radiation wavelength are held constant, a plot of concentration versus absorbance should be a straight line if Beer's law holds. Two conditions are implied in the derivation of Beer's law. The first is that the radiation being measured is monochromatic. In an IR spectrometer, especially under low-resolution conditions, what is actually measured is the intensity of a narrow region of the spectrum that may be significantly wide compared with the width of the absorption band being measured. This means that the absorbance deduced is an average absorbance for a finite wavelength section of the absorption band. This can cause deviation from Beer's law. The second condition for linearity is that the sample absorptivity not change with concentration. If the concentration range is large, the sample environment may change. For example, when the concentration is low, the sample is surrounded by solvent, whereas when the concentration is high, the sample is surrounded by other sample molecules. The change in environment can cause absorptivity changes and deviations from Beer's law. When the concentration change causes hydrogen-bonding changes, deviations can be severe. If Beer's law does not hold exactly, a plot of concentration versus absorbance (for constant cell thickness and wavelength) will not be a straight line, but will have a slight curvature. For narrow concentration ranges, such a plot will be nearly linear.

C. Measurement of Absorbance

There are a variety of ways of measuring the absorbance. In a single-beam spectrum the vertical coordinate I is a measure of the source radiation intensity (at a given wavelength) attenuated by sample absorption and atmospheric carbon dioxide and water vapor absorption. A reference single-beam spectrum is run with the same instrument conditions but without the sample. Let this vertical coordinate be I_0. If the vertical coordinates of these two spectra are divided, wavelength by wavelength, a spectrum is generated where the vertical coordinate is transmittance or, when multiplied by 100, percent transmittance. This is not the percent transmittance of the chemical sample, but rather that of the whole cell assembly. In addition to the absorption of radiation by the sample, radiation can be lost by cell reflection and scattering of radiation, as well as by beam blockage by an undersized cell aperture or by beam attenuation. A background correction must be made.

If the vertical coordinate of the spectrum is linear with percent transmittance, then the percent transmittance reading at, say, the bottom of a sample band is taken to be proportional to I. The percent transmittance at the same wavelength is read for the background point, that is, where the recorder pen would be if there were no sample band. This is taken as proportional to I_0. From these values for I and I_0 the background corrected sample absorbance can be calculated as $\log_{10}(I_0/I)$.

Sometimes the vertical coordinate of the spectrum is the absorbance value rather than the percent transmittance. In this case, the absorbance reading at the band peak is corrected by subtracting the absorbance reading of the background point at the same wavelength where the recorder pen would be if there were no sample band.

If the solvent has some absorption at the analytical wavelength or wave number, both the sample and solvent contribute to the total absorbance. Beer's law is additive; that is, the total absorbance is equal to the sum of the abc values for each component. The solvent should absorb less strongly than any of the solvent–sample mixtures at the analytical wavelength. In this case, the pure solvent is used for a "zero sample" or background reading as before. The absorbance of pure solvent is subtracted from the absorbance of the solution of the sample plus solvent. Because Beer's law is additive, this absorbance difference should be proportional to sample concentration when the same cell and instrument settings are used if Beer's law holds.

D. Integrated Band Intensities

Absorbance values of band peaks are most commonly used for quantitative analysis. However, a peak height measurement is somewhat sensitive to instrument resolution. If the resolution is reduced, a narrow-shaped band will become broader and the peak height will be reduced. The integrated intensity is a measure of the total band area, and this shows much less sensitivity to instrumental resolution. Furthermore, it has greater theoretical significance in that the integrated band intensity is a measure of the total radiation energy absorbed for this vibrational mode and it is proportional to the square of the change in dipole moment with respect to the change in the normal coordinate.

One way to characterize the band intensity is to express it as the integrated absorptivity as a function of wave number. From Beer's law the absorptivity a is equal to $A/(bc)$, which means that the integrated intensity can be expressed as the absorbance A, integrated over the whole band, divided by bc. If the spectrum of a solution is plotted with the horizontal coordinate linear with wave number $\bar{\nu}$ (cm^{-1}) and the vertical coordinate linear with absorbance A, then the band area, corrected for background and divided by bc, is the integrated intensity. If the cell length is measured in centimeters and the concentration is measured in moles per cubic centimeter, then the integrated absorptivity is in centimeters per mole. Other units have been used.

E. Base Line Method

A popular way to do a quantitative analysis from recorded spectra is the base line method. In this method, taking a percent transmittance or absorbance reading of the band peak is straightforward. It is the method of getting the background percent transmittance or absorbance that gives the method its name. Again, one tries to imagine where the recorder pen would be if the component being measured were not present. If one has an isolated band in a region with no other absorption, the background line is easy to draw as a line tangent to the spectral background. If one has a band that comes on the sloping side of a band of the solvent or of a major component, as in Fig. 15, one has a choice. One could try to sketch a rounded background

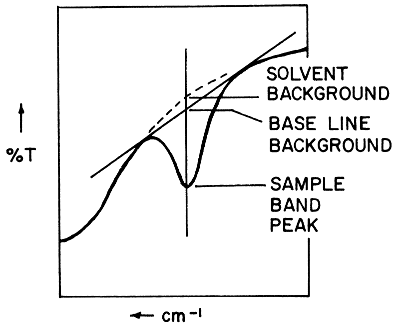


FIGURE 15 Part of an infrared spectrum showing the base line construction for the base line method used in quantitative analysis.

for the "zero sample" condition and use that for the background point. This can be used but is somewhat subjective and may not be very reproducible. An alternative is to use a line tangent to the spectrum at the band wings or even connecting some more distant points. The more closely the line approaches the true background, the better, but it is not necessary that the tangent line and the true background coincide exactly. A consistent base line construction should yield a reproducible measured absorbance for a given concentration, and the absorbance should vary linearly with concentration. Again, this is calibrated with standards. One of the advantages of the method is that it can reduce or eliminate a correction for the finite background absorption of the solvent or major component. Some care should be taken in choosing the tangent points for the base line. These points should not be too sensitive to concentration variations of other components that may be present.

F. Ratio Methods

In most solid-state spectra, the sample film thickness and the amount of sample in a KBr disk or a Nujol mull are not known. It is still possible to do quantitative analysis by the ratio method. This is based on the fact that, in a given mixture, the absorbance ratio of any two bands in one spectrum should be independent of the sample thickness. Consider a two-component mixture, each component of which has an analytical band with no interference from the other component. The absorbance ratio for the two analytical bands is, from Beer's law,

$$\frac{A_1}{A_2} = \frac{a_1 b_1 c_1}{a_2 b_2 c_2} \quad \text{or} \quad \frac{A_1}{A_2} = \frac{a_1 c_1}{a_2 c_2}. \tag{19}$$

The b values are identical and cancel since the absorbances are measured for the same sample preparation. Since a_1/a_2 is a constant, the absorbance ratio is proportional to the concentration ratio. One can also write

$$c_1 = \frac{100}{1 + c_2/c_1} \tag{20}$$

which, when multiplied out simply states that the sum of the two concentrations c_1 and c_2 is 100%. In this equation the ratio c_2/c_1 is replaced by its equivalent from the previous equation to give

$$c_1 = \frac{100}{1 + (a_1/a_2)(A_2/A_1)}. \tag{21}$$

The a_1/a_2 ratio is a constant that can be determined by measuring A_2/A_1 for one standard of known concentration. Once a_1/a_2 is evaluated, the concentration c_1 of an unknown can be calculated from the absorbance ratio.

Another ratio method is the internal standard method. In this method a known amount of an internal standard material is added to the sample mixture. The internal standard is chosen so that it has no absorption at the analytical wavelength for the sample and vice versa. Again from Beer's Law, the absorbance ratio is

$$\frac{A}{A_s} = \frac{ac}{a_s c_s} \quad \text{and} \quad c = c_s \frac{a_s A}{a A_s} \tag{22}$$

where the subscript s denotes the internal standard. Here the internal standard concentration c_s is known, and the constant ratio a_s/a can be determined from one sample of known concentration, after which the concentration can be calculated from the absorbance ratio.

G. Gas State Analysis

In a quantitative analysis involving gases, the concentration term c in Beer's law is replaced by the pressure or partial pressure p of the gas being measured. Beer's law reads

$$A = abp. \tag{23}$$

In this case, the measurement of the background absorbance is usually that of the evacuated cell or the cell filled with nonabsorbing gas such as nitrogen. There is a complication in the use of Beer's law in the vapor phase that is called pressure broadening. Not only is the absorbance of a gaseous component a function of its partial pressure; it is also a function of the total pressure. This means that the absorbance of a gaseous component with a fixed partial pressure can be changed by introducing another completely nonabsorbing gas such as nitrogen. The rotational fine structures of gas-phase bands are broadened by collisions between the molecules of the component being measured, and other gas molecules and collisions vary in frequency and severity as the pressure increases. For this reason the total pressure is often kept constant in a quantitative analysis of gases. The desired partial pressure of the gas to be analyzed is introduced into an evacuated cell. Then the total pressure is increased to a standard value such as 760 mm Hg by introducing nitrogen, for example, before measuring the absorbance.

H. Multicomponent Analysis

In the most general case for quantitative analysis, one wishes to measure several components of the mixture and there are no isolated wave numbers. This means that, at the best analytical wave number for one component, the other components have finite absorption that interferes with the measurement. Fortunately, Beer's law is additive, which means that at any given wave number in the

spectrum, the total absorbance is equal to the sum of the *abc* values for each component. One can specify that all the measurements will be done in the same cell, so the thickness is constant for the whole analysis. This means that *b*, the cell thickness, can be combined with a, the absorbtivity, to give a new constant *k*, which replaces *ab* in Beer's law. If there are three components, 1, 2, and 3, absorbing at a specific wave number, then Beer's law is

$$A = k_1c_1 + k_2c_2 + k_3c_3. \tag{24}$$

In order to measure these three concentrations in an unknown, three absorbances are needed at three different wave numbers.

$$A_1 = k_{11}c_1 + k_{12}c_2 + k_{13}c_3 \tag{25a}$$

$$A_2 = k_{21}c_1 + k_{22}c_2 + k_{23}c_3 \tag{25b}$$

$$A_3 = k_{31}c_1 + k_{32}c_2 + k_{33}c_3. \tag{25c}$$

Here, A_1, A_2 and A_3 absorbances at wave numbers 1, 2, and 3, where concentrations of components 1, 2, and 3 are best measured. In k_{12} for example, the first subscript is for wave number 1 and the second subscript is for the concentration of component 2. Three standards (std.) are prepared with known concentrations with suitable ranges for the analysis. These are all run at the first analytical wave number to give three A_1 equations.

$$\text{(std. 1)} \quad A_{11} = k_{11}c_{11} + k_{12}c_{21} + k_{13}c_{31} \tag{26a}$$

$$\text{(std. 2)} \quad A_{12} = k_{11}c_{12} + k_{12}c_{22} + k_{13}c_{32} \tag{26b}$$

$$\text{(std. 3)} \quad A_{13} = k_{11}c_{13} + k_{12}c_{23} + k_{13}c_{33}. \tag{26c}$$

Here, the second subscript on *A* and *c* is for the standard number. For these three equations, where the *A* values and the nine concentrations (C_{11} etc.) in the standards are known, the three unknown *k* values can be evaluated. This same procedure is used for the A_2 and A_3 equations (25b and c) to evaluate all nine of the unknown *k* values. Once these are known, the three equations (25a, b, and c) can be used to calculate all the unknown concentrations (c_1, c_2, and c_3) in a sample from the measured absorbencies. This is called the *method of simultaneous equations*. While this method is straightforward, the disadvantages are that Beer's law nonlinearities are difficult to handle and all the components in the mixture must be accounted for. Also, only a limited number of analytical wave numbers can be used.

For use in repetitive analyses, these quantitative methods have been computerized. In the example discussed, the absorbances are expressed as functions of the various concentrations. A computerized version of this is called the *classical least squares* (CLS) (or the Kmatrix method), as it gives the least squares prediction for the concentrations. This method can use many more than the minimum number of analytical wave numbers, or even all of them. This full-spectrum method improves the precision over those that use only a few wave numbers. Corrections can be added for Beer's law deviations or fitting spectral baselines. However, all components present must be included in the calibration mixtures.

In a variation of this type of analysis, the concentrations are expressed as functions of the various absorbances rather that vice-versa as before. This is called the *inverse least squares* (ILS) (or the Pmatrix method). An advantage of this method is that a quantitative analysis can be performed on some components using calibrated standards, even if some other components with unknown concentrations are present in the standards in amounts bracketing those in the samples. A disadvantage is that it is not a full-spectrum method. In the analysis, there must be at least as many standards for calibration as there are analytical wave numbers used.

Two factor analysis methods that are used are the principal components regression (PCR) and the partial least squares (PLS). In the PRC method, the concentrations are expressed as functions of the principal components (PC) instead of absorbances as in ILS. The PC are orthogonal vectors that are linear combinations of the original spectral data of the standards. Here, PC1 accounts for the maximum variability in the data, and PC2 accounts for the maximum variability not accounted for by PC1, etc. The other method PLS, is similar to the PCR method except that the PCs are weighted. The weighting is based on the correlation of the PCs with concentration. These are full-spectrum methods like CLS, but like ILS, one can analyze one component at a time. These methods are most often used for quantitative analysis in the near infrared region because of the broadness and overlapping nature of the bands here.

VI. GROUP FREQUENCIES

A. Concept of Group Frequencies

Bands at certain frequencies in the IR spectra have been related to the presence of certain functional groups in the chemical structures. For example, in the spectra of a series of unconjugated ketones, a band common to all is a strong band near 1715 cm^{-1}. This has been assigned to the stretching of the carbony1 bond and is a group frequency for unconjugated ketones. There are many other bands in these ketone spectra, which differ from molecule to molecule, especially below 1300 cm^{-1}. These are fingerprint-type bands that can be used to distinguish one ketone from another. A large body of empirical knowledge has been built up about the characteristic group frequencies, which has proved to be very useful to the chemist. The vibrations that give rise to group frequencies

are those whose vibrational forms are nearly the same in a series of related molecules.

B. Vibrational Interaction

If all the bonds in a molecule vibrated separately, the diatomic vibrational frequency formula given earlier could be used to predict the whole spectrum. The fact is that bonds do not usually vibrate separately and much interaction occurs. However, interactions occur only if vibrations have the same type of symmetry. For example, in a planar molecule such as vinyl fluoride, planar vibrations do not interact with non-planar vibrations since these have different symmetries with respect to the plane.

One type of interaction occurs when two identical bonds share a common atom. An example is the H_2O molecule, which has two OH bonds with a common oxygen atom. In the H_2O molecule, one OH bond cannot be vibrationally excited without also exciting the other identical OH bond at the same time. The second bond vibrates either in phase or out of phase with the first. In one case, both bonds stretch at the same time and contract at the same time, and in the other, one bond stretches while the other bond contracts. In the H_2O gas-phase spectra, these are observed at 3652 and 3756 cm^{-1}, respectively. Thus, the in-phase and out-of-phase vibrations do not have the same frequencies. The main reason for this is that, when the two vibrating bonds are not at the equilibrium length, both bonds exert restoring forces on the common oxygen atom. The force resultant is different for the in-phase and out-of-phase stretch vibrations, and this affects the frequencies. This is called vibrational interaction.

There are several XY_2 groups that have correlatable stretching frequencies. For example, alkane CH_2 groups have stretching vibrations that absorb near 2930 and 2850 cm^{-1}, amino NH_2 groups have stretching vibrations that absorb near 3370 and 3300 cm^{-1}, and sulfone SO_2 groups have stretching vibrations that absorb near 1300 and 1130 cm^{-1}. In each case the higher wave number involves the out-of-phase stretch and the lower wave number involves the inphase stretch.

Consider the XYZ group where the X—Y bond and the Y—Z bond have quite *different* frequencies when unconnected. Examples include the C—O—H group in alcohols, the C—C≡N group in nitriles, and the C—S—H group in mercaptans. Here the other groups on the carbon are ignored. The XYZ group as a whole has two stretching frequencies, but the interaction is different from that for the XY_2 case. In the high-frequency vibration of the XYZ group, only the atoms of the high-frequency bond move appreciably. In the examples given, this means that when the high-frequency OH, C≡N, or SH bonds vibrate, the attached carbon hardly moves since it is attached to the high-frequency bond with a low-frequency C—O, C—C, or C—S bond. This means that the rest of the molecule does not affect the OH, C≡N, or SH vibrations mechanically very much since the nearly stationary attached carbon atom localizes the vibration. In this manner, one can see that such group vibrations as OH, C≡N, and SH are group frequencies; that is, their presence in a molecule gives rise to absorption in a predictable frequency range.

While the OH vibration is mechanically unaffected by the rest of the molecule, the OH group has a force constant that can be changed by hydrogen-bonding effects. An alcohol in dilute CCl_4 solution has a free OH band near 3640 cm^{-1}. Hydrogen bonding lowers the OH frequency and increases the bandwidth and intensity. A pure alcohol is hydrogen bonded (OH $\cdots$ O) and absorbs broadly and strongly near 3300 cm^{-1}.

In a ketone, the high-frequency C=O bond is connected to the rest of the molecule by two low-frequency C—C bonds. In the carbonyl vibration the two attached carbons hardly move, making the carbonyl a good group frequency, mechanically nearly independent of the rest of the molecule. The carbonyl frequency can be varied by electron donation or withdrawal effects from the attached groups, and these can shift the frequency. These effects are fairly well understood, which means that the shifts are predictable.

The CH_2, NH_2, and SO_2 groups discussed earlier are attached to the molecule by low-frequency C—C, C—N, or C—S bonds, so the attached atoms hardly move, isolating the CH_2, NH_2, and SO_2 stretching vibrations. This means that both the in-phase and out-of-phase stretching vibrations for these groups are good group frequencies, nearly independent mechanically of the rest of the molecule.

In a group such as the C=S group, the vibration is not isolated like a C=O vibration. The C=S is attached to the rest of the molecule by C—C or C—N bonds, which have nearly the *same* frequency as the C=S bond. As a result, interaction will take place and more than one vibration will involve C=S stretching.

In such groups as CH_3 and SO_3, there are three identical bonds. These interact, so the group vibrates as a whole in three different modes: an in-phase stretch and two different out-of-phase stretch vibrations. These are good group frequencies because the group is connected to the molecule with a low-frequency C—C or C—S bond. In a group such as a benzene ring there are six identical or nearly identical C—C bonds, which interact to give six different stretching modes. Some of these are group frequencies.

C. Survey of Group Frequencies

Table IV contains some selected group frequencies used in qualitative analysis. Some discussion of the data follows. In the region from 4000 to 2000 cm^{-1}, various XH groups absorb. In the region from 3700 to 3100 cm^{-1}, OH groups

TABLE IV Selected Spectra Structure Correlations

Group[a]	**Wave number region (cm^{-1})**[b]	**Assignment**[c]
Alkanes		
$R—CH_2—R$	2936–2916 s	Out-of-phase str.
$R—CH_2—R$	2863–2843 m	In-phase str.
$R—CH_2—R$	1475–1450 m	CH_2 def.
$—CH_2—CH_2—CH_2—CH_2—$	726–722 w	In-phase rock[d]
$R—CH_3$	2972–2952 s	Out-of-phase str.
$R—CH_3$	2882–2862 m	In-phase str.
$R—CH_3$	1475–1450 m	Out-of-phase def.
$R—CH_3$	1383–1377 m	In-phase def.
$C(CH_3)_2$	1389–1381 m	Sym. in-phase def.
$C(CH_3)_2$	1372–1368 m	Antisym. in-phase def.
$C(CH_3)_3$	1401–1393 m	Sym. in-phase def.
$C(CH_3)_3$	1374–1366 s	Antisym. in-phase def.
Olefins		
C═C	1680–1630 mw	C═C str.
$C═CH_2$	3100–3075 w	CH_2 out-of-phase str.
$R—CH═CH_2$	995–985 s	Trans CH wag[e]
$R—CH═CH_2$	910–905 s	$═CH_2$ wag
$R_2C═CH_2$	895–885 s	$═CH_2$ wag
RCH═CHR (trans)	980–965 s	Trans CH wag
RCH═CHR (cis)	730–650 m	Cis CH wag
$CH_2═CH—CO—OR$	970–960 m	$═CH_2$ wag
$CH_2═CH—O—R$	820–810 s	$═CH_2$ wag
X≡Y and X═Y═Z		
C—C≡C—H	2140–2100 w	C≡C str.
C—C≡C—H	3340–3267 s	C—H str.
C—C≡N	2260–2240 m	C≡N str.
—C≡N (conjugated)	2240–2220 v	C≡N str.
S—C≡N	2170–2135 m	C≡N str.
$C═C≡CH_2$	2000–1900 s	Out-of-phase str.
—N═C═O	2275–2263 s	Out-of-phase str.
—N═C═S	2150–2050 s	Out-of-phase str.
Aromatics		
Aromatic CH	3100–3000 w	CH str.
Aromatic ring	1620–1585 v	Ring str.
Aromatic ring	1590–1565 v	Ring str.
Aromatic ring	1525–1470 v	Ring str.
Mono and meta	710–665 s	Ring bend out of plane
Five adjacent ring H's	800–730 s	In-phase CH wag
Four adjacent ring H's	805–735 s	In-phase CH wag
Three adjacent ring H's	825–750 s	In-phase CH wag
Two adjacent ring H's	860–795 s	In-phase CH wag
Isolated ring H	935–810 s	CH wag
Carbonyls		
R—CO—R	1725–1705 s	C═O str.
Conjugated ketones	1700–1640 s	C═O str.
R—CO—H	1740–1720 s	C═O str.
Aryl—CO—H	1710–1685 s	C═O str.
R—CO—O—R	1750–1735 s	C═O str.

continues

TABLE IV (*continued*)

Group[a]	Wave number region (cm^{-1})[b]	Assignment[c]
Conjugated esters	1735–1715 s	C═O str.
Lactone (γ)	1795–1740 s	C═O str.
—COOH (dimer)	1720–1680 s	C═O str.
—$CO_2^-Na^+$	1650–1540 s	Out-of-phase str.
—$CO_2^-Na^+$	1450–1360 m	In-phase str.
R—CO—N	1690–1630 s	C═O str.
O—CO—N	1740–1683 s	C═O str.
R—CO—CI	1810–1795 s	C═O str.
Anhydride (noncyclic)	1825–1770 s	In-phase C═O str.
Anhydride (noncyclic)	1755–1715 ms	Out-of-phase str.
Anhydride (cyclic)	1870–1845 m	In-phase C═O str.
Anhydride (cyclic)	1800–1750 s	Out-of-phase str.
Carbonyl substituents		
CH_2—C═O	1440–1405 m	CH_2 def.
CH_3—C═O	1375–1350 ms	CH_3 in-phase def.
H—CO—R	2900–2800 w	CH str.[f]
H—CO—R	2775–2695 w	CH str.[f]
H—CO—R	1420–1370 w	CH def.
—CO—O—R	1300–1150 s	C—O str.
—CO—OH (dimer)	3000 broad m	OH str.
—CO—OH	1315–1280 s	C—O str.
—CO—NH_2	3520–3180 s	NH_2 str. (two bands)
—CO—NH_2	1635–1600 m	NH_2 def.
—CO—NH—(noncyclic)	3470–3250 m	NH str.
—CO—NH—(noncyclic)	1550–1510 m	NH def. + C—N stretch
Alcohols and ethers		
C—OH (unbonded)	3641–3593 w	OH str.
C—OH (H-bonded)	3500–3000 s	OH str.
CH_2—OH	1075–1000 s	C—O str.
R_2CH—OH	1150–1075 m	C—O str.
R_3C—OH	1210–1100 m	C—O str.
Aryl—OH	1260–1180 s	C—O str.
CH_2—O—CH_2	1140–1085 s	C—O—C out-of-phase str.
Aryl—O—CH_2	1310–1210 s	Aryl—O str.
Aryl—O—CH_2	1050–1010 m	O—CH_2 str.
Nitrogen groups		
—NH_2	3550–3330 mw	NH_2 out-of-phase str.
—NH_2	3450–3250 mm	NH_2 in-phase str.
—NH_2	1650–1590 mw	NH_2 def.
CH_2—NH—CH_2	1146–1132 m	C—N—C out-of-phase str.
Aryl—N	1330–1260 m	Aryl—N str.
C═N	1690–1630 m	C═N str.
CH_2—NO_2	1556–1545 s	NO_2 out-of-phase str.
CH_2—NO_2	1388–1368 m	NO_2 in-phase str.
Aryl—NO_2	1530–1500 s	NO_2 out-of-phase str.
Aryl—NO_2	1370–1330 w	NO_2 in-phase str.
Chlorine		
C—CI	830–560 m	C—CI str.
C—CH_2—CH_2—CI (trans)	730–723 m	C—CI str.

continues

TABLE IV (*continued*)

Group[a]	Wave number region (cm^{-1})[b]	Assignment[c]
C—CH_2—CH_2—Cl (gauche)	649–635 m	C—Cl str.
CH_2—Cl	1300–1240 m	CH_2 wag
Sulfur		
—SH	2590–2540 w	SH str.
CH_2—S	1270–1220 m	CH_2 wag
—SO_2—	1400–1300 s	SO_2 out-of-phase str.
—SO_2—	1200–1100 s	SO_2 in-phase str.
—SO_3Na	1230–1120 s	SO_3 out-of-phase str.
—SO_3Na	1080–1025 m	SO_3 in-phase str.
Phosphorus		
PH	2440–2275 m	PH str.
P=O	1320–1140 s	P=O str.
P—O—C	1050–970 s	P—O—C str.
Silicon		
SiH	2250–2100 s	SiH str.
Si—CH_3	1280–1255 s	CH_3 in-phase def.
Si—O—R	1110–1000 s	Si—O—C str.
Si—O—Si	1110–1000 s	Si—O—Si str.
Boron		
BH	2640–2350 s	BH str.
B···H···B	2220–1540 ms	BH str.
B—O	1380–1310 s	BO str.

[a] R is an alkane group.
[b] Wave number regions given in cm^{-1}. Relative intensities are indicated by s (strong), m (medium), w (weak), and v (variable).
[c] Abbreviations: str., stretch; def., deformation; sym., symmetric; antisym., antisymmetric.
[d] CH_2 rock is CH_2 rotation in the CH_2 plane.
[e] Wag is rotation out of the CH_2 plane or out of the aryl or olefinic plane.
[f] Fermi resonance doublet, CH stretch + CH deformation overtone.

in alcohols and phenols and water have bands. Various types of NH also absorb here, as well as acetylenic CH groups. In the region from 3100 to 3000 cm^{-1}, absorption occurs for aryl CH and olefinic CH. In the region from 3000 to 2800 cm^{-1}, various types of CH_3 and CH_2 groups absorb. Strongly hydrogen bonded hydrogens in acidic compounds absorb broadly in the region from 3100 to 2400 cm^{-1}. Finally, SH, BH, PH, and SiH groups absorb in the region from 2600 to 2100 cm^{-1}.

Various types of triple bonds such as nitriles and cumulated double bonds such as isocyanate groups absorb in the region from 2300 to 1900 cm^{-1}. Various types of double bonds, including C=O, C=N, and C=C absorb from 1900 to 1550 cm^{-1}. Aromatic ring vibrations absorb near 1600 and 1500 cm^{-1}.

Hydrogen bending vibrations absorb from 1600 to 700 cm^{-1}. Bending vibrations for CH_2 and CH_3 absorb from 1500 to 1350 cm^{-1}, and very useful out-of-plane CH wag vibrations in olefins and aromatics absorb from 1000 to 700 cm^{-1}. In olefins these can be used to distinguish vinyls, vinylidines, and cis and trans 1,2-disubstituted olefins. In aromatics, these bands can usually distinguish the number of adjacent aromatic hydrogens between substituents: five for monosubstituted, four for ortho, three and one for meta, and two for para.

In the range from 1400 to 900 cm^{-1} vibrations from SO_2, SO_3, P=O, and many types of C—O groups have strong bands. Usually below 1350 cm^{-1} only strong bands are interpreted since there are many bands in this region that are not group frequencies. This, of course, is only a very small sample of the large body of group frequency correlations available for qualitative analysis that can be found in IR texts.

SEE ALSO THE FOLLOWING ARTICLES

Infrared Astronomy • Infrared Technology • Macromolecules, Structure • Microwave Molecular Spectroscopy • Radiometry and Photometry • Raman Spectroscopy

BIBLIOGRAPHY

Chalmers, J. M., and Dent, G. (1997). "Industrial Analysis with Vibrational Spectroscopy," Royal Society of Chemistry, Analytical Spectroscopy Monographs, Cambridge CB4WF, UK.

Coleman, P. A. ed. (1993). "Practical Sampling Techniques for Infrared Analysis," CRC Press, Boca Raton, FL.

Colthup, N. B., Daly, L. H., and Wiberley, S. E. (1990). "Introduction to Infrared and Raman Spectroscopy," 3rd ed. Academic Press, New York, NY.

Crocombe, R. A., Olson, M. L., and Hill, S. L. (1987). "Computerized Quantitative Infrared Analysis," pp. 95–130. ASTM STP 934. McClure, ed., American Society for Testing and Materials, Philadelphia, PA.

Diem, M. (1993). "Introduction to Modern Vibrational Spectroscopy," Wiley—Interscience, New York.

Griffiths, P. R., and de Haseth, J. A. (1986). "Fourier Transform Infrared Spectrometry," Wiley, New York, NY.

Lin-Vien, D., Colthup, N. B., Fately, W. G., and Grasselli, J. G. (1991). "The Handbook of Infrared and Raman Characteristic Frequencies of Organic Molecules," Academic Press, New York, NY.

Messerschmidt, R. G., and Harthcock, M. A. (1988). "Infrared Microspectroscopy," Marcell Dekker Inc., New York, NY.

Pouchert, C. J. (1985). "The Aldrich Library of FT-IR Spectra," Vols. 1 and 2, The Aldrich Chemical Company, Inc., Milwaukee, WI.

Liquid Chromatography

Neil D. Danielson
Miami University

I. Analytical HPLC
II. Small- and Large-Scale HPLC
III. Separation Techniques
IV. Development of Separation Strategy
V. Sample Derivatization Chemistry

GLOSSARY

Bonded-phase packing Stationary support such as silica having chemically attached functional groups.

Capillary electrochromatography A hybrid of capillary electrophoresis and HPLC involving mobile phase transport by electrosmotic flow not by pump pressure through a packed capillary column.

Dead volume Any volume between the injector and detector such as in connecting tubing and fittings where separation does not occur.

Elution Removal of sample components from a column.

Gradient elution Changing the mobile phase composition with time at a specified rate during the chromatographic separation.

Guard column Short column attached in front of the separation column to adsorb compounds that could degrade the latter column.

Ion chromatography Rapid and efficient ion-exchange separation of ions usually with conductivity detection.

Ion-exchange Type of liquid chromatography for the separation of ions based on their charge affinity for the ionic functional groups of the column packing.

Ion-exclusion Type of liquid chromatography that separates neutral compounds by partitioning into the liquid in between and inside the ion exchange beads as well as adsorption on the polymer backbone.

Ion-pair chromatography Reversed-phase separation of charged compounds by addition of an oppositely charged reagent to the mobile phase.

Normal-phase Type of liquid chromatography in which the stationary phase is polar (e.g., silica) and the mobile phase is nonpolar (e.g., hexane).

Pellicular packing Column packing consisting of solid glass beads enveloped with a 1–2-μm porous layer.

Reverse(d)-phase Type of liquid chromatography in which the stationary phase is nonpolar (e.g., C-18 hydrocarbon) and the mobile phase is polar (e.g., water–methanol).

Size exclusion Type of liquid chromatography that separates sample components by molecular size generally in order from increasing to decreasing molecular weight.

Solute The separated species on analyte.

THE SEPARATION of compounds or ions as they pass down a column due to the differing distribution of the sample components between the liquid mobile phase and

a particle-supported stationary phase is the technique termed liquid chromatography (LC). There are two major types of LC, classical and high-performance liquid chromatography (HPLC). Classical LC uses large columns approximately 2 × 50 cm, packed with large porous particles 50–250 μm in diameter. Sample volumes in the milliliter range are often required. The mobile phase is generally gravity-fed at slow flow rates because the deep pores of the packing limit mass transfer and separation times can be on the order of hours. Fraction collection of the components for later spectroscopic identification is common. Although inefficient and not sensitive, classical LC requires no special equipment and is often adequate for some organic synthesis and biochemical research. The evolution of HPLC occurred in the late 1960s as the technology became available to manufacture pellicular (see Glossary for definition) and 10-μm size porous silica particles. These tiny particles when packed in smaller columns of 0.4 × 25 cm require pumping of the mobile phase to overcome the pressure drop of 1000 to 3000 psi. Because of the shallower pores in these packing, mass transfer of the analyte from the bulk of the mobile phase to the surface of the stationary phase is faster, permitting separation times on the order of 10 to 30 min with good resolution. Because smaller sample volumes of usually 20 μl are necessary to not exceed the capacity of these columns, detectors equipped with flow cells have been developed for more sensitive and convenient detection of the separated components. This description of HPLC is considered to be analytical in nature and will be the major emphasis of this article. However, preparative and microbore HPLC which bracket analytical HPLC with respect to flow rates, sample size, and column dimensions will be briefly discussed. Capillary HPLC and the relatively new technique, capillary electrochromatography, will also be compared.

I. ANALYTICAL HPLC

A block diagram for a typical instrument or high-performance liquid chromatograph is shown in Fig. 1. Each part of the chromatograph will be described in some detail with respect to analytical HPLC first.

A. Solvent Mixing

The mobile phase reservoirs are generally screw-top plastic-coated glass bottles with inlet lines to allow for degassing of the solvents for 10–15 min either by vacuum or sparging with helium or both. A positive pressure of helium should be maintained in the solvent bottles. Degassing is important to prevent gas bubbles from lodging in the pump, column, or the detector. On-line degassing units are commercially available. Two micrometer inlet filters are attached to the liquid-carrying Teflon tubing to prevent any insoluble particles from being drawn to the pump where they could cause check valve malfunction. As shown in Fig. 1, blending of the mobile phase constituents in the proper ratio is accomplished by the solenoid proportioning valves A, B, and C. By controlling the fraction of time each valve is opened using the computer system through a digital-to-analog (DAC) interface, a reproducible mobile phase made up of two or three components can be formed in the mixing chamber. The volume of the mixing chamber should be kept small to minimize gradient lag time at the column. The mixing chamber, often magnetically stirred, contains a 2 μm frit for final mobile phase filtration before it reaches the pump.

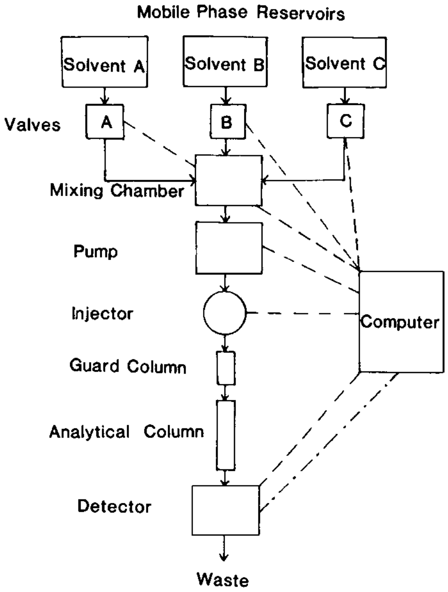


FIGURE 1 Block diagram of a high-performance liquid chromatograph; —, Mobile phase flow; ---, computer control of device; —·—· data acquisition and readout.

High-pressure mixing is an alternative method for generation of the desired mobile phase. The outlets of two or more HPLC pumps are connected into a mixing chamber. Basically, the solvents are mixed on the high-pressure side of the pumps. The composition of the desired mobile phase is dependent on the selected pumping speeds of each pump, which are often computer controlled. The resultant flow rate represents the combined output of the individual pump flow rates. Although precise control of the gradient composition is possible using this system, the expense of multiple pumps is a disadvantage.

B. Pump

The HPLC pump must meet several criteria. These are (1) pumping capability up to 6000 psi, (2) pulseless

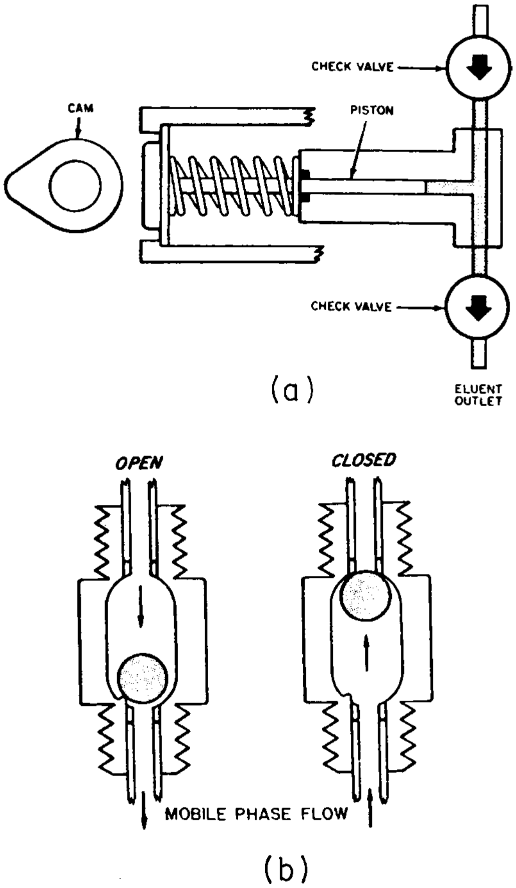


FIGURE 2 A single-piston reciprocating pump. (a) schematic diagram (b) operation of a check valve. [From Yost, R. W., Ettre, L. S., and Conlon, R. D. (1980). "Practical Liquid Chromatography," Perkin–Elmer Corporation, Norwalk, Connecticut, p. 146. Reprinted with permission.]

solution output, (3) a flow rate range of 0.1 to 10 ml/min with 0.5% relative reproducibility, and (4) noncorrosive wetted parts. Although a variety of pumps such as the syringe and pneumatic types have been used, the reciprocating pump comes closest to fulfilling the above requirements at a modest cost. A schematic diagram of a simple reciprocating pump is shown in Fig. 2a. Basically, the operation involves a motor rotating a cam that pushes a sapphire piston back and forth in a small chamber. Ruby ball check valves (Fig. 2b) permit solvent flow in only one direction. The pumping rate is generally adjusted by controlling the speed of the motor. The flow rate is often computer controlled through a DAC and is generally set between 0.5 and 3 ml/min. Advantages of this pump include good flow-rate reproducibility and compatibility with gradient mobile phases. The primary disadvantage is the definite pulselike nature of the delivered solution stream that can cause detector baseline noise at sensitive settings. To alleviate this problem, dual-head and more recently triple-head reciprocating pumps have been developed so that the mobile phase is always being pumped out while the other piston(s) are refilling. A pulse dampener is usually not required for these pumps unless very sensitive detector settings are used. Commercial pulse dampers such as a coiled flattened tube or a diaphragm unit are available. They add dead volume to the system, however, which is undesirable when changing solvents. Some pumps have titanium not stainless steel all wetted parts for better buffer salt compatibility. Most pumps permit external washing of the pump heads without disassembly to eliminate salt buildup.

C. Injector

In liquid chromatography, introduction of the sample ideally on the column or very close to it is important to minimize sample diffusion and band broadening. Sample sizes are small, ~5–100 μl, with 20 μl being a common size. Most HPLC injectors used today are loop valves as shown in Fig. 3. A syringe is used to overfill a sample loop while the mobile phase or eluent is bypassed directly to the column (Fig. 3a). Upon switching the valve, the eluent is diverted, displacing the sample from the loop and onto the column (Fig. 3b). Since the valve is designed to withstand 6000 psi, the mobile phase flow need not be stopped. Convenience and good reproducibility are the primary advantages. Error values less than 0.5% are possible. The rotor seal is prone to scratching from particulates and prefiltration or centrifugation of the sample should be carried out if deemed necessary. Seal wear can generate small black particles which need to be taken out of the mobile phase before they reach the column using a low dead volume on-line filter placed after the injection valve.

D. Column Dimensions and Hardware

Most analytical HPLC columns are made of precision, smooth-bore stainless steel tubing. Uniformity of the inside walls is important to minimize channeling effects at the wall/packing interface. Dimensions for stainless

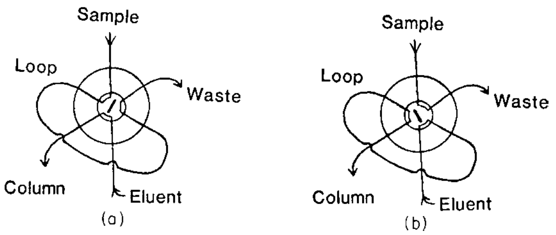


FIGURE 3 A rotary sample valve. (a) Valve position for filling sample loop and (b) for introduction of sample into the column.

steel columns are generally 5–30 cm in length with a 2.3–6.2-mm inside diameter. The length of the column is often determined by the particle size of the packing since pressure drop is inversely proportional to the square of the particle diameter (dp). The pressure drop (Δp) of a packed bed can be more completely described as

$$\Delta p = \eta L\, \mu\phi/dp^2,$$

where η = mobile phase viscosity, L = column length, μ = linear velocity, and ϕ = flow resistance factor ($\approx$500). Small diameter columns can have better peak sensitivity (see Section I.E) but sacrifice sample capacity.

Stainless steel zero dead volume fittings such as a $\frac{1}{16}$ in. $\rightarrow$ $\frac{1}{16}$ in. union (Fig. 4a) are often required to make connections. The $\frac{1}{4}$ in. $\rightarrow$ $\frac{1}{16}$ in. reducing union fittings (Fig. 4b) are used to connect the column to the injector and detector. Both types of fittings can easily withstand 6000 psi pressure. Pressed in the column fittings are 0.2-μm frits to contain the packing particles in the tube and prevent any particulates from reaching and disturbing the packing bed. The connecting volume between the column-injector and column-detector must be kept at a minimum because bandbroadening (variance or σ^2) is directly proportional to the fourth power of the tubing radius but only proportional to the column length, and inversely proportional to the flowrate. Therefore, $\frac{1}{16}$ in. stainless steel tubing with a 0.01 in. or smaller inside diameter is used for all fitting connections. Recently, metal free column hardware and polyethyletherketone (PEEK) tubing have become available to design HPLC instruments particularly suited for separations of biological samples.

E. Column Packing Supports

Column packing supports for HPLC can be divided into two main classes, silica and polymer types. Silica (Fig. 5a) has several desirable properties such as excellent pressure stability, high surface area, good control of size and shape, and ease of availability. Of the silica packings, pellicular and porous are the two major kinds. Pellicular packings, developed first as an alternative to classical large particle LC supports, basically consist of a solid glass bead on which a 1–2 μm porous layer of silica has been deposited. Interaction of the sample components occurs only in this thin layer. Originally 40-μm size particles were made but now smaller particles are available. Pellicular particles, fairly inexpensive in price, are sometimes used in guard columns. Guard columns, usually 3–5 cm in length, protect the analytical column from irreversibly adsorbed sample constituents often found in complex biological matrices. In addition, a silica guard column can protect a silica analytical column from dissolution due to a high pH mobile phase by saturating the mobile phase with dissolved silica. Porous silica particles that are 10, 5, or 3 μm in size are more commonly packed in analytical columns. Recently, columns packed with 2-μm particles have been reported in the literature. The 3–10-μm size particles can be commercially obtained either irregular or spherical in shape with a surface area ranging from 50 to 500 m^2/g.

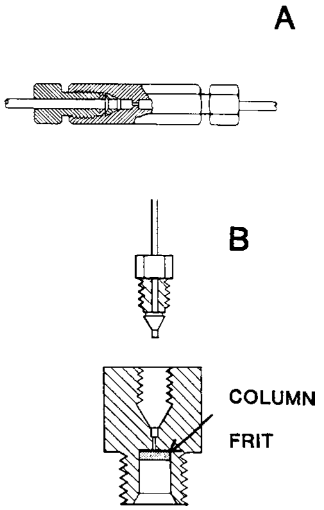


FIGURE 4 HPLC zero dead volume $\frac{1}{16}$ in. $\rightarrow$ $\frac{1}{16}$ in. connecting (A) and $\frac{1}{4}$ in. $\rightarrow$ $\frac{1}{16}$ in. reducing (B) unions. [From S. Schram (1980). "The LDC Basic Book on Liquid Chromatography," Milton Roy Co., p. 79. Reprinted with permission from Milton Roy, Inc., and R. W. Yost, L. S. Ettre, and R. D. Conlon (1980). "Practical Liquid Chromatography," Perkin–Elmer Corporation, Norwalk, Connecticut, p. 245.]

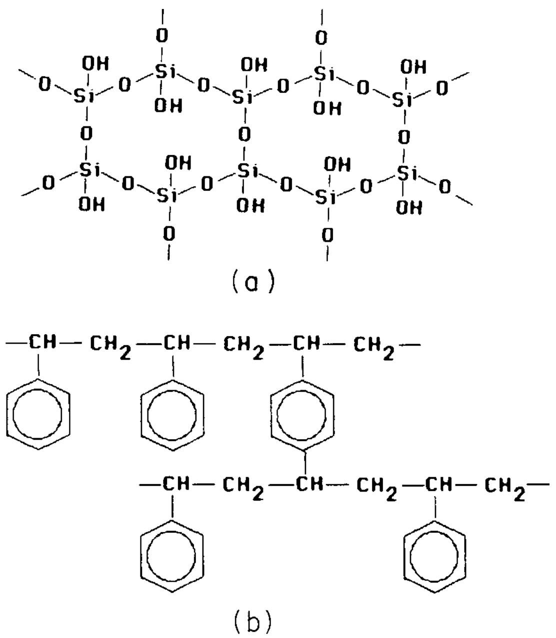


FIGURE 5 Chemical structures of silica (a) and polystyrenedivinylbenzene (PSDVB) (b).

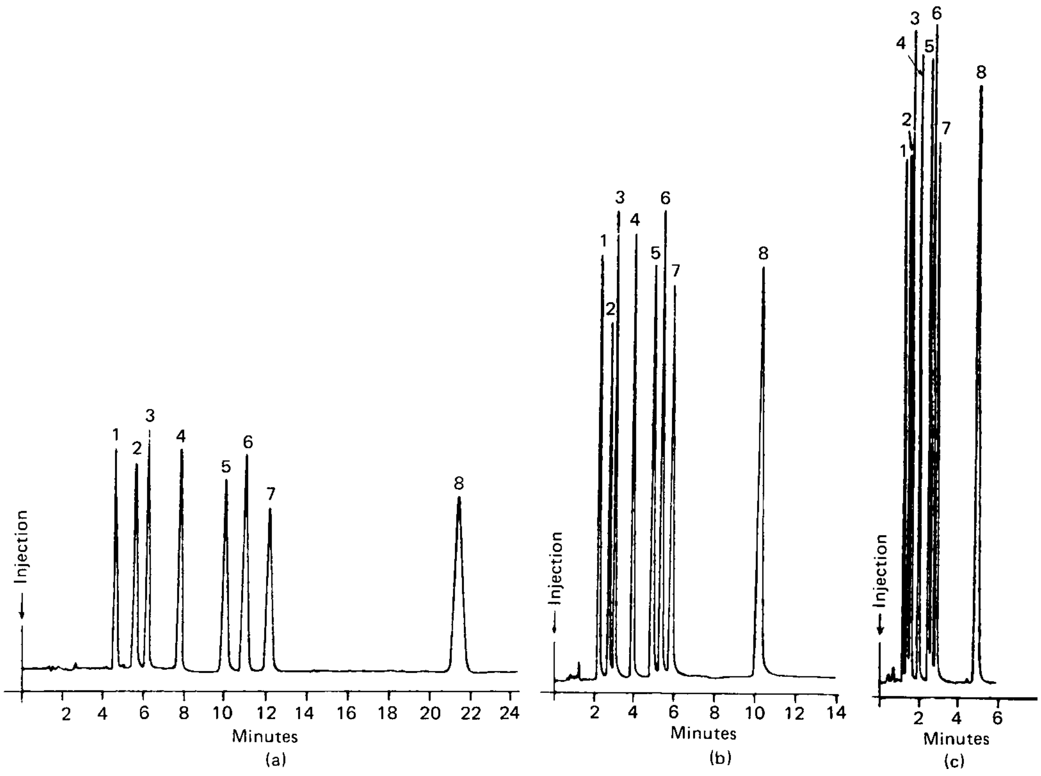


FIGURE 6 Effects of particle size on column performance: (a) 10 μm, (b) 5 μm, and (c) 3 μm. Ultrasphere C-18 columns: (a) 30 cm × 4.6 mm, (b) 15 × 4.6 mm, and (c) 7.5 × 4.6 mm. Mobile phase: 60–40 methanol–water. Flow rate = 1 ml/min. Temperature = 30°C. Pressure: (a) 810 psi, (b) 1600 psi, and (c) 2250 psi. Peaks: (1) Phenol, (2) Benzaldehyde, (3) Acetophenone, (4) Nitrobenzene, (5) Methylbenzoate, (6) Anisole, (7) Benzene, and (8) Toluene. [Reprinted with permission from Beckman/Altex Scientific.]

The pore size of the silica particles must be large enough to permit easy entrance and exit of the sample molecules. However, since pore size is inversely proportional to the surface area of the packing, the pores should not be excessively larger than the sample components of interest. For relatively small organic or inorganic molecules, a pore size of 60 Å is sufficient. For the separation of large molecules such as polymers or proteins, a pore size of at least 250 Å is preferred.

Packing materials for the HPLC separation of biomolecules that have not only 500–1500 Å pores but also a network of 6000–8000 Å transecting tunnels have been developed by Regnier. These highly porous materials, when packed into columns, permit mobile phase velocities 2–5 times higher than those for conventional wide pore silica columns. For the rapid reversed-phase separation of peptides and proteins, both small 2-μm porous wide pore (200 Å) silica and 2-μm pellicular silica microspheres packed in 3-cm columns have been developed. For a five component mixture of proteins, separation times of less than a minute were possible. Nonporous monodisperse 1.5-μm silica beads developed by Unger have been shown to be effective as column packings for the reversed-phase separation of proteins and are commercially available.

As can be seen in Fig. 6, the advantages of columns packed with smaller particles are faster analysis times, improved solute sensitivity, and decreased solvent consumption. It can be shown that the peak height maximum ($C_{\max}$) can be calculated from the following equation, $C_{\max} = (C_s V_s / V_r)(N/2\pi)^{0.5}$, where C_s and V_s are the concentration and volume of sample injected respectively, V_r is the retention volume, and N = number of theoretical plates. The number of theoretical plates can be calculated easily from the chromatographic data as $N = 16(t_r/w_b)^2$, where t_r = peak retention time and w_b = width of the peak at baseline in time units. Because V_r is proportional to the volume of the column ($\pi r^2 \times L$, where r = column radius and L = column length), $C_{\max}$ will increase proportionally to the ratio of the square of the radii as column radius decreases. Going from a column of 4.6 mm to one of 2.1 mm and assuming all other variables are constant, the improvement in $C_{\max}$ can be predicted to be almost 5 times. An efficiency comparison of different octyldecyl (C-18) modified silica columns using plate count (N) is shown

TABLE I Quantitative Column Comparison at Optimum Velocity[a]

Column type	N/column	N/m	N/sec	N/psi
Conventional 5-μm C-18 silica (250 × 4.6 mm)	21800	87200	36.8	12.5
Microbore 10-μm C-18 silica (500 × 1.0 mm)	19100	38200	9.05	63.7
Small particle 3-μm C-18 silica (100 × 4.6 mm)	11900	119000	99.2	3.81

[a] From McCoy, R. W., and Pauls, R. E. (1982). *J. Liq. Chromatogr.* **5,** 1869. Reprinted with permission from Marcel Dekker, Inc.

in Table I. As predicted by the Van Deemter equation, the number of plates per meter increases as particle diameter decreases. The Van Deemter equation can be expressed as

$$\text{HETP} = A + Bu + C_\text{m}u + C_\text{s}u,$$

where HETP = height equivalent of a theoretical plate, u = mobile phase velocity, A = eddy diffusion term dependent on the particle diameter of the packing (d_p), B = longitudinal diffusion term dependent on the diffusion constant of the analyte in the mobile phase (D_m), C_m = the mass transfer term involving the solute from the bulk mobile phase to the surface of the stationary phase and is dependent on d_p^2/D_m, and C_s is the mass transfer of the solute into and out of the stationary phase and is dependent on the square of the film thickness (d_f) of the stationary phase. A good rule of thumb is $N \approx 3000\,L/dp$, where L = column length (in centimeters) and dp = particle diameter (in micrometers). The small particle column was lowest in plates generated per unit pressure, indicating high-column back pressure could be a problem with some mobile phases.

Silica can be used as is for normal phase HPLC in which the mobile phase is nonpolar in nature, such as hexane/chloroform and the column-packing surface is polar. For reversed-phase HPLC in which the mobile phase is polar, such as water or methanol, and the column-packing surface is nonpolar, the silica must be chemically altered. Silanization reactions are carried out to covalently attach long chain hydrocarbon groups to the silica surface. Further details of the importance of silica supports are provided in Section III, "Separation Techniques."

Porous zirconia particles coated with polybutadiene make a reversed-phase HPLC column packing that offers both excellent pressure stability and chemical inertness throughout the entire pH range. Because of the complex surface chemistry of zirconia, mobile phase additives such as phosphate or fluoride are added to facilitate the separation of acidic compounds such as anti-inflammatory drugs (aspirin, acetoaminophen, and ibuprofen) or basic drugs such as antihistamines. Zirconia columns provide higher plate counts than polymeric columns and comparable efficiencies to silica based columns. Titania has also been investigated as a stable HPLC column packing but is less well studied.

Polymeric packings for HPLC have been developed largely to overcome the mobile-phase pH constraints of silica. In general, for routine use of silica-based packings, the pH of the mobile phase must be maintained between pH 2 and 7.5. This is because the silica itself can dissolve at alkaline pH, and cleavage of the siloxane bonds holding the functional groups occurs at acid pH. Another advantage of polymeric packings is that few residual polar groups, such as nonreacted hydroxyls found in silica, are present that can cause peak tailing through hydrogen bonding with certain solutes. Most polymer HPLC packings are polystyrene-divinylbenzene (PS–DVB) resins cross-linked at 10% or greater to ensure sufficient particle rigidity (Fig. 5). Even so, the column pressure limit for most polymers is limited to about 3000–5000 psi. Most PS–DVB HPLC packings are spherical in shape and 5 or 10 μm in size. The pore size and surface area of PS–DVB particles can also be controlled similar to the range of values given for silica. However, generally polymeric packings are less efficient (20,000–50,000 plates/m) than silica packings. Because of the aromatic (moderately nonpolar) nature of PS–DVB, it is used primarily for reversed-phase HPLC unless chemically modified. Functionalization of PS–DVB resins for use as ion-exchange packings is the other dominant use of these supports. Other polymers such as polyvinylpyridine, polyacrylamide, polyvinylalcohol, and various floropolymers such as derivatized polychlorotrifluoroethylene have also been used as reversed-phase packings. Further details of the importance of polymeric supports to ion-exchange chromatography in particular are provided in Section III.

F. Column Packing Technique

Both the dry-fill and wet-fill packing methods have been used to prepare HPLC columns. The dry-fill approach involves vertical tapping of the column with simultaneous rapping along its side to cause good consolidation of the column bed. This procedure is recommended for the packing of rigid particles with a diameter greater than 20 μm, such as pellicular spheres. Small porous particles have high surface energies with respect to their mass and clump together when attempted to be packed dry. Preparative HPLC columns are often made using this "tap-filled" method.

The wet-fill or slurry packing techniques are recommended for particles less than 20 μm in diameter. In this approach, the slurry is pumped quickly under high pressure into the column blank. The solvent chosen to prepare the slurry must wet the particles and keep them well dispersed. In addition, it is desirable to choose a solvent with a density that approaches the particle density, the so-called balanced density situation. Although suspected cancer causing agents such as tetrabromoethane or chloroform can be used, this method is recommended primarily for 10 μm particles. Particles 5 μm or less settle very slowly, and other solvents should be considered. Methanol or isopropyl alcohol are commonly considered for silica packings or polymers. For small reversed-phase silica particles, acetone-hexane mixtures, because of their low viscosity, have been employed. Because of the possibility of charging nonpolar reversed-phase packings due to the rapid liquid flow from the packing pump, 80% methanol–20% sodium acetate (0.1%) in water has been recommended.

A high-pressure pneumatic pump capable of being pressurized to 8,000 to 10,000 psi by closing a shutoff valve is generally necessary to pack efficient columns. Upon opening the valve, the released pressure forces the packing slurry from a stainless steel reservoir into the column blank in a few seconds. This "slamming" process generally will permit a uniform packed bed of particles in the column. The column bed is considered stable if the liquid flow from the column is constant after several repressurization cycles. After no liquid is observed to drip from the column, it is carefully removed from the packing reservoir. The column ends are cleaned off with a razor blade to permit the fittings to be attached. The column should be checked with a test mixture and mobile phase chosen to give peak retention times from 5 to 15 min. A column plate height two to three times the particle diameter should be obtainable if the column is well packed. Peaks should be well shaped with no shoulders or split tops. This is a sign of channeling and the column must be repacked. Excessive peak width is usually a sign of a void volume and again the column should be repacked.

G. HPLC Detectors

A wide variety of HPLC detectors have been developed to try to fill both the high-sensitivity and universal-detection requirements. Some of these are adaptations of well-known GC detectors such as flame ionization and electron capture. This discussion will focus primarily on the design and appropriate applications of the seven most common and commercially available HPLC detectors. A listing of these detectors is shown in Table II. The HPLC detectors can generally be classified as either responsive to a change in the property of the mobile phase when a solute (sample component) is present or to a property of the actual solute itself.

An example of the latter is the most widely used HPLC detector, the UV–VIS spectrophotometer (Fig. 7). Its popularity is due to a wide range of applicability, excellent stability, and low cost. This detector is designed to be equipped with a low-volume (8 μl or less) flow cell usually 1 cm in pathlength. The low volume of the cell is important to minimize bandbroadening, as is the radius of the connecting tubing to the column. Lenses are used to focus as much light as possible through the 1- or 2-mm diameter cell. Most filter HPLC spectrophotometers are equipped with a mercury lamp. This source emits 90% of its radiation at 254 nm, which is an excellent wavelength for the detection of aromatic compounds. Different filters and/or special phosphor converters placed after the lamp allow other wavelengths to be selected. The zinc lamp is used for shorter wavelengths such as 214 nm. Usually the light is focused through dual flow cells, one considered the sample side and the other the reference side. If the mobile phase

TABLE II Characteristics of Liquid Chromatography Detectors[a]

Detector basis	Type[b]	Temperature sensitive?	Linear range	Noise level[c]	Maximum[d] sensitivity
UV absorption	S	Low	5×10^3	2×10^{-4} AU	2×10^{-10}
Refractive index	G	$\pm 10^{-4}$°C	5×10^5	2×10^{-6} RIU	1×10^{-7}
Fluorometry	S	Low	6.4×10^3	0.005 v	10^{-11}
Electrochemical	S	1.5%/°C	5×10^3	2×10^{-9} μamp	10^{-12}
Conductometric	S	2%/°C	1×10^6	0.05 μMho	10^{-8}
IR absorption	S	Low	10^3	0.01 AU	10^{-6}
Mass spectrometry	G	None	10^3	—	10^{-10}

[a] Most of these data were taken from Snyder, L. R. and Kirkland, J. J. (1979). "Introduction to Modern Liquid Chromatography, 2nd ed." Wiley-Interscience, New York, p. 162. Reprinted by permission of John Wiley & Sons, Inc.

[b] G = General; S = selective.

[c] AU = Absorbance units; RIU = refractive index units.

[d] Sensitivity for a favorable sample in g/ml.

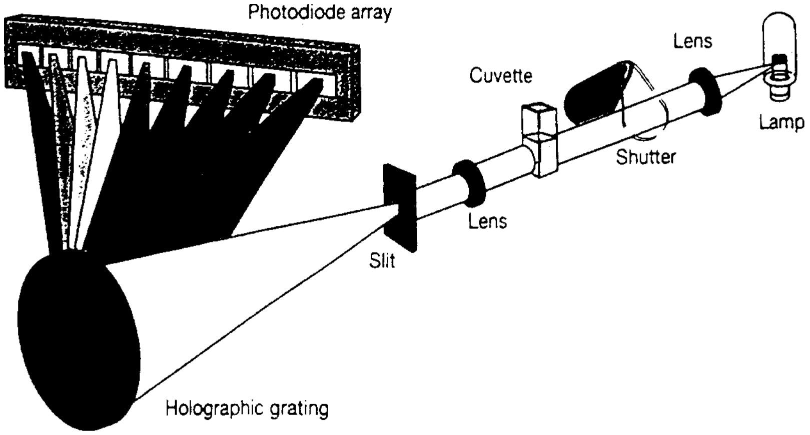


FIGURE 7 Diagram of a photodiode array instrument. For HPLC, the cuvette is replaced by a flow cell similar to that used for the standard UV–VIS detector. [From Siouffi, A-M., Chapter 1, in "Food Analysis by HPLC." (L. M. L. Nollet, ed.), Marcel Dekker, New York.

does not absorb, which is often the case, the reference side is usually air. However, occasions arise where it is advantageous to pump the mobile phase through both cells for subtraction purposes. The detector is quite stable to temperature fluctuations; however, air bubbles that lodge or pass through the flow cell will cause baseline spikes. A UV–VIS HPLC detector equipped with a monochromator for specific wavelength selectivity is more versatile but also more expensive. Deuterium and tungsten lamps are used to provide ultraviolet and visible light, respectively. Often multiple wavelengths can be monitored simultaneously. The photodiode array (PDA) UV–VIS detector has now become quite common as the standard HPLC detector ordered with a new instrument. The PDA detector employs a reverse optics design which allows the entire spectrum of light to be dispersed onto a diode array (see Fig. 7). Each diode of the array is responsible for detecting light of a narrow wavelength range depending on the desired spectral range to be covered and the speed of data acquisition. This instrument can take a spectrum over a preselected wavelength range in a fraction of a second as the solute elutes from the column. Therefore, qualitative (peak purity) as well as quantitative information about the unknown sample component can be obtained.

The differential refractometer is perhaps the second most widely used HPLC detector because of its universal nature. Potentially, any substance with a refractive index (RI) different from the mobile phase is detectable. The deflection type RI detector consists of a light source passing two beams of monochromatic light through a double prism that constitutes the sample cell and reference cell. If the mobile phase composition changes, the altered refractive index causes the beam to be deflected from its initial position on the photomultiplier detector. The electrical signal produced is proportional to the light position, which is a function of the solute concentration. Although the RI detector is versatile, it is not particularly sensitive (microgram level) and is very prone to temperature fluctuations. Careful temperature control of the detector cell with a water jacket is crucial for maintaining a stable baseline. Modifications in the electronic design and the use of a laser source have also improved the detectability of the RI detector.

Luminescence detectors can be either based on fluorescence or chemiluminescence. The fluorescence HPLC detector is basically a fluorometer equipped with a flow cell slightly larger in volume than the UV–VIS detector to permit more fluorescent light from the solute to reach the photomultiplier. Again, the primary modification is the use of lenses to focus the excitation source light onto the flow cell. Filter instruments are often more sensitive than monochromator instruments because of greater light throughput. Again, lasers have allowed the use of small volume flow cells with lower detection limits. The fluorescence detector, although useful for only a select class of aromatic hydrocarbons or derivatives, is about 100 times more sensitive than a UV–VIS detector. For example, polyaromatic hydrocarbons can be determined at the ppb level. Signal stability with respect to temperature is good. The coupling of fluorescence to HPLC has generally minimized background signals from impurities and oxygen quenching effects. If the lamp of the fluorescence detector is turned off, it can operate as an effective chemiluminescence detector. Postcolumn addition of reagents such as luminol and a metal catalyst for the detection of an oxidizing agent such as H_2O_2 is necessary.

The electrochemical (EC) HPLC detector is basically a small electrode composed usually of glassy carbon mounted in a flow cell (Fig. 8a). The auxiliary electrode is

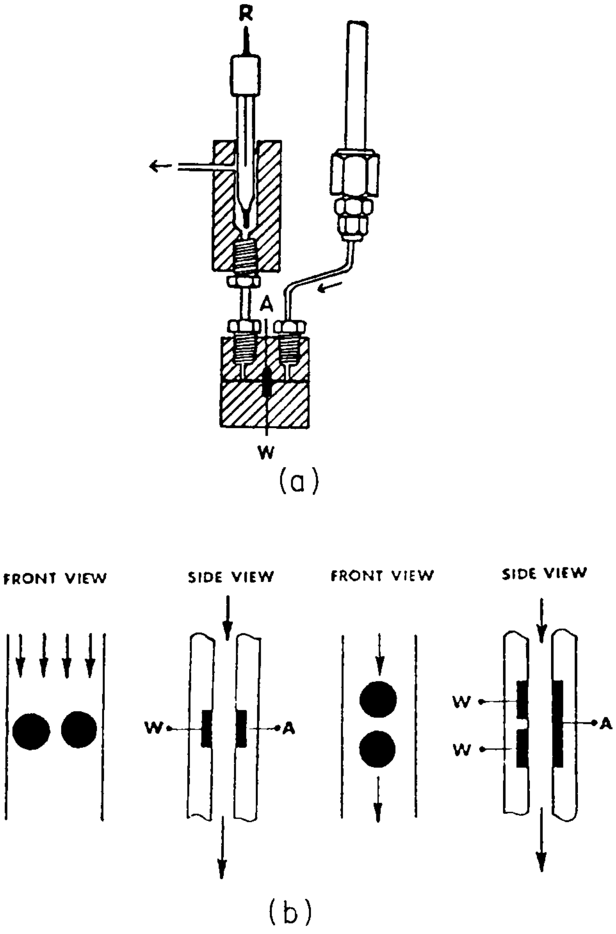


FIGURE 8 Electrochemical detector cell (Bioanalytical Systems, Inc.). (a) Diagram of the flow cell. A = auxiliary electrode, W = working electrode. R = reference electrode. (b) Dual thin-layer working electrodes in parallel (1) and series (2) configurations. [From Bratin K., and Kissinger, P. T. (1981). *J. Liq. Chromator.* **4,** 321–57. Reprinted with permission from Marcel Dekker, Inc.]

opposite the carbon working electrode while the Ag/AgCl reference electrode is slightly downstream. Upon application of a voltage to the cell, oxidation of the solute of interest occurs and the resultant current is measured. Oxidation of organic compounds such as phenols to the corresponding quinone is the preferred mode of operation. Reduction because of the interference of dissolved oxygen is more difficult but has been shown to be useful for nitro compounds. Dual electrode flow cells with either a parallel or series arrangement (Fig. 8b) have been developed. The parallel arrangement (1) permits oxidation or reduction of the separated components at two different electrode potentials providing extended detection capabilities. The series arrangement (2) is useful to remove dissoloved oxygen by reduction at the first electrode before detection of the sample solutes at the second electrode. The coulometric EC detector which can oxidize or reduce a major fraction of the eluting peak gives a stable response with low detection limit. Only reversed phase HPLC using a buffered mobile phase is compatible with EC detection because the solution must be electrically conductive. However, this is not considered to be a practical limitation because about 80% of the HPLC applications are of the reversed-phase type. Because nanoamp current levels can be measured, the EC detector is very sensitive, on par with the fluorescence detector. HPLC with EC detection is the method of choice for the determination of catecholamines and neurotransmitters in biological fluids.

The pulsed amperometric detector (PAD) developed by Johnson and co-workers using an Au or Pt electrode has permitted the direct detection of aliphatic alcohols including carbohydrates, amines, and sulfur compounds. Fouling of these electrodes is prevented by application of both positive (to eliminate sample adsorption) and negative (to reduce any metal oxide) reactivation step potentials on the order of 100 ms before resetting the potential for detection of the analyte. The analytical current is usually sampled near the end of the detection potential pulse to permit decay of the charging current. The oxidation of these aliphatic compounds such as carbohydrates is facilitated in basic solution at about pH 12, so postcolumn addition of 0.1 *M* NaOH or the use of a polymeric column with a basic mobile phase is required. Detection limits of alcohols and carbohydrates are at the 10 ppb level. Alkanolamines, amino acids, and sulfur compounds other than sulfonic acids and sulfones can also be detected.

The conductivity detector is based on the ability of ions to conduct electricity across two electrodes in a flow cell between which an electric field is applied. The current measures is proportional to the conductivity of the solution. Usually a sinusoidal wave potential is applied to the electrodes and the only current measured is that in phase with the applied potential. Most conductivity detectors for HPLC can be set electronically to compensate for the background eluent conductivity. In addition, because the mobility of ions varies with solution temperature, a thermistor is mounted close to the cell to permit electronic compensation. The conductivity detector has proved to be important for ion-exchange HPLC of simple inorganic and organic ions.

The evaporative light-scattering detector is an excellent alternative to the refractive-index detector for nonvolatile organic compounds with no UV–VIS chromophore because of the improved detection limits. Operation can be described by a three-step process (see Fig. 9) involving nebulization of the solvent, evaporation of the solvent in a heated tube to give particles of pure solute, and finally measurement of the scattered laser light caused by the solute particles. Detection limits can be improved by generating large particles through efficient nebulization and

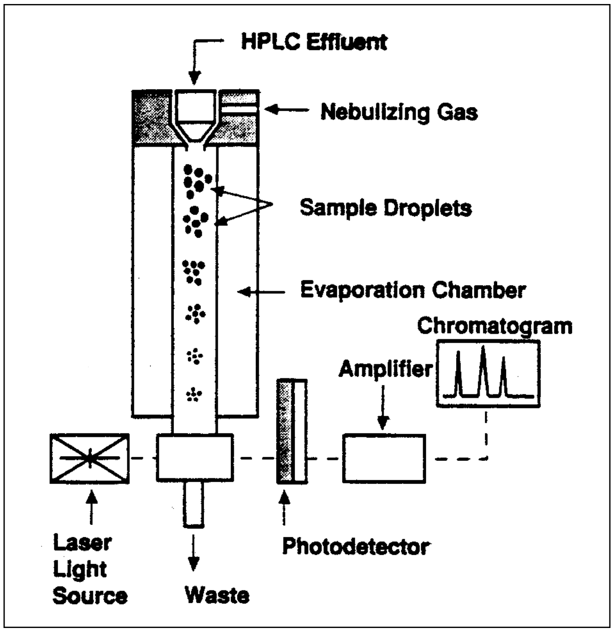


FIGURE 9 Diagram of an evaporative light-scattering detector [Reprinted by permission from Alltech Associates.]

temperature control of the evaporation tube. Minimizing particulate impurities in the mobile phase is also important. Common reversed-phase organic solvents and water up to 25% even with volatile buffers can be effectively removed. Typical applications include lipids, sugars, and triglycerides.

Mass spectrometry (MS) has now become a user-friendly detector for LC not only because of the increased reliability but also the lower cost. The most common commercial LC/MS instrument is based on the electrospray interface (Fig. 10). The mobile phase containing the separated sample components is converted to an aerosol of electrically charged droplets by means of strong shear forces of the nebulizing gas and the strong electrostatic field (2–6 kV) in the spray chamber. Ions of one polarity are attracted to the droplet surface by the electrostatic field causing the droplets to disperse in a fine spray or electrospray. For example, if the solution contains acetic acid and the nebulizer needle is held at a positive potential, positively charged sample molecular ions (MH^+) will be formed. Because the sample solution is not heated when the aerosol is created, thermal decomposition of the analytes is not a problem. Before the ions can be mass analyzed, the solvent must be removed to yield the base ion. A counterflow of heated nitrogen gas causes the desolvation process forcing the like-charged surface ions closer together until a Coulombic explosion produces tiny droplets. Continued evaporation of the remaining solvent is rapid until base analyte ions are produced that can then enter the metallic opositely charged end of the glass capillary leading to the mass spectrometer. The nebulizer and capillary are set at right angles to each other to minimize mobile phase or sample matrix contamination of the mass spectrometer. Because the glass capillary separates the atmospheric pressure region of the source from the vacuum region of the mass spectrometer, the analyte ions are pushed by a pressure gradient through the capillary as a supersonic jet flow. As the ions exit, they are electrostatically repelled to the first skimmer by the like-charged metallic end of

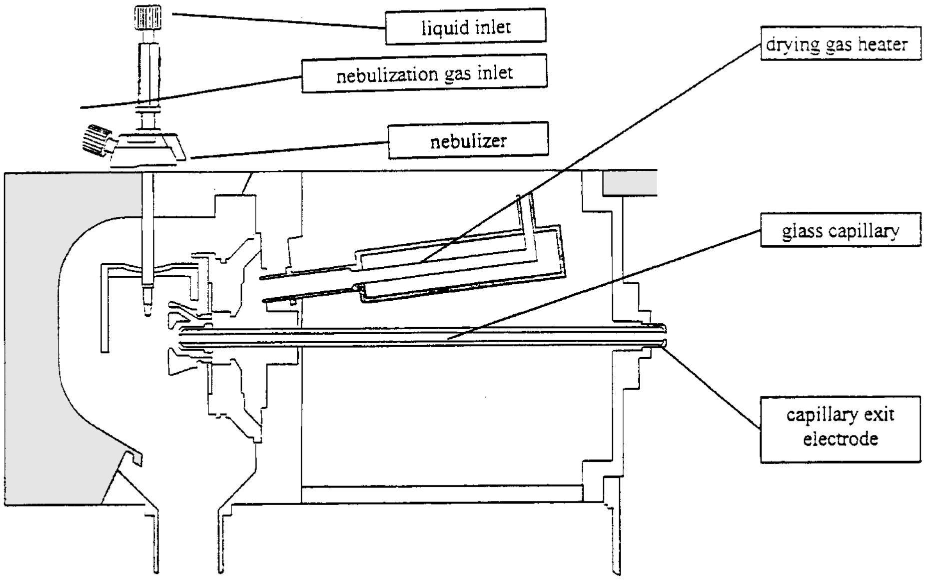


FIGURE 10 Diagram of the Bruker orthogonal electrospray LC/MS interface. [Reprinted by permission from the Bruker-HP Esquire LC Operations Manual, Version 3.1.]

the capillary. In the low–pressure region, skimmers, octopoles, and lenses concentrate and focus the analyte ions into the ion trap which is an ion storage and mass analysis device. The ion trap permits MS/MS, and additional molecular-structure information of MS^n can also be obtained. The electrospray interface accommodates a broad range of mobile phase compositions including volatile buffers and 100% water at flow rates from 1–1000 μl/min. A mass range up to 6000 m/z permits the analysis of some protein and polymer samples.

On-line infrared and NMR detection are both possible in conjunction with analytical or preparative HPLC. Despite the strong infrared absorbance of water and organic solvents, the cylindrical internal reflection (CIRCLE) cell with a cell volume of 24 μl offers a short pathlength allowing for this background absorbance to be ratioed out. Detection limits with the CIRCLE cell are in the low micromolar range. Many NMR instruments can be fitted with a flow-through probe with a cell volume of about 120 μl. The stopped-flow mode may be required to allow for signal averaging to minimize the background due to undeuterated solvents.

A summary of the important parameters of the HPLC detectors previously discussed is shown in Table II. The UV–VIS or PDA detector should be present on all HPLC instruments for general application use. The other detectors should be added as demand for identification and analysis of certain classes of compounds arises.

II. SMALL- AND LARGE-SCALE HPLC

The previous instrumental description is typical for analytical HPLC that can handle injected samples between 5 and 100 μl. However, sometimes it is of interest for either very small or very large samples to be separated. Microbore and capillary HPLC can respectively analyze small and even smaller samples while semipreparative and preparative HPLC can respectively analyze large and larger samples.

A. Microbore HPLC

The typical microbore HPLC column is $\frac{1}{16}$-in. tubing with an inside diameter of 1 to 2 mm and a length ranging from 25 to 100 cm. Particle size of the packing material and column-packing techniques are similar to those previously described. Because of the narrow column diameters, and longer lengths, flow rates on the order of 10 to 50 μl/min are the norm. As shown in Table I, the plate count per unit time and length are somewhat lower than with analytical columns. However, these plates were achieved with a minimal sacrifice in pressure drop as indicated by the N/psi value. Because of the reduced amount of packing material, sample sizes on the order of 0.2 to 1 μl are used. To prevent peak broadening, the volume of the flow cell must be less than 2 μl and the dead volume caused by fittings must be essentially eliminated. The advantages of microbore HPLC are (1) a major decrease in solvent consumption and hence cost, permitting more exotic solvents to be used, (2) a greater sensitivity when limited to a small sample, and (3) a potential for greater separation efficiencies because of longer columns. A typical chromatogram is shown in Fig. 11.

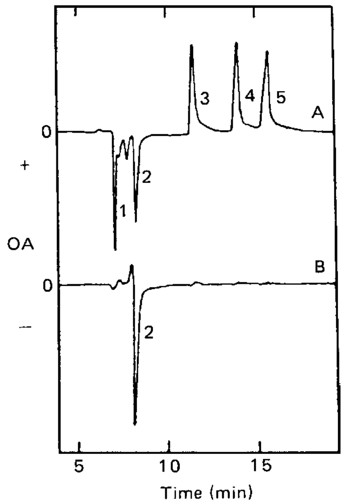


FIGURE 11 Microbore chromatography with polarimetry detection (OA = optical activity). Chromatogram of a mixture containing (1) injection peak, (2) *l*–2 octanol, (3) decane, (4) tetradecane, and (5) hexadecane. Mobile phase: (A) (–)-2methyl–1–butanol in CH_3CN (50 : 50) and (B) (±)-2methyl–1–butanol in CH_3CN (50 : 50). Column = 1 mm × 25 cm 5-μm C-18 silica. Sample size = 0.5 μl; Flow rate = 20 μl/min. [From Bobbitt, D. R., and Yeung, E. S. (1984). *Anal. Chem.*, **56,** 1577. Reprinted with permission by the American Chemical Society.]

B. Capillary HPLC

Capillary LC has become more widely accepted as commercial equipment to accommodate the low flow rates of 1–5 μl/min, sample injection sizes of 60 nl, and the capillary detector flow cells has become available. Columns generally 100–350 μm ID × 25 cm in length packed with 3 or 5 μm particles are also commercially available. The main advantages of capillary LC are the small sample size and improved sensitivity as compared to analytical or microbore HPLC. Using the equation described in Section I.D, a 320-μm capillary could theoretically provide 200 times improvement in sensitivity as compared to a standard 4.6-mm ID column assuming the same sample size could be injected. However, for large volume injections with capillary LC (see Fig. 12), an on-column focusing

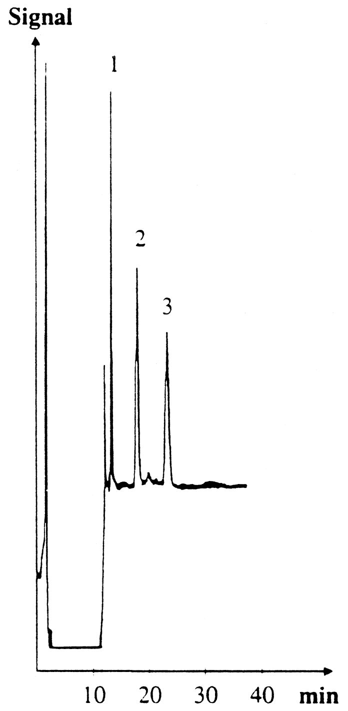


FIGURE 12 On-column focusing large volume injection capillary LC-UV separation of retinoids. The injection volume was 100 μl, the operating temperature 50°C, and the volumetric flow rate 20 μl/min during sample introduction. The concentration of each retinoid was 50 ng/ml: (1) all-*trans*-retinol, (2) 13-*cis*-retinoic acid, (3) all-*trans*-retinoic acid. [From Molander, P., Gunderson, T. E., Haas, C., Greibrokk, T., Blomhoff, R., and Lundanes, E. J. (1999). *Chromatogr. A* **847,** 59–68.]

method must be used employing a noneluting mobile phase at a higher flow rate and temperature to reduce backpressure. Once the sample has been loaded, a valve must be switched to permit the eluting mobile phase to start the separation process. Capillary LC because of the low flow rates is particularly well suited for mass spectrometry detection. The reduced mobile phase flow rate leads to a decrease in spray droplet size with an increase in electrospray ionization efficiency. Because mass spectrometry is a concentration sensitive technique, the sensitivity advantage as previously mentioned in comparison to standard analytical columns is also important.

A relatively new technique called capillary electrochromatography (CEC) is a hybrid of capillary electrophoresis (CE) and capillary LC. CEC is potentially more versatile than CE or HPLC because separation is based on both mobility differences (if the compounds are charged) and reversed-phase retention (effective for neutral organics and charged compounds with hydrophobic moieties). Mobile-phase transport through a typical 50–200 μm ID capillary packed with 3-μm particles is achieved by electroosmotic flow (EOF) instead of a pressure gradient as in HPLC. The origin of EOF is the electrical double layer that is formed at the solid–liquid interface due to the negatively charged silanols of the capillary wall and the positively charged cations in solution such as Na^+. Upon application of an electric field across the capillary, the positive ions are attracted to the negative detector end of the capillary and move the bulk flow by viscous drag. The advantage of EOF over conventional pressure driven flow is that no column backpressure is generated and the flow profile is pluglike not laminar in nature. Because of the pluglike flow profile, eddy diffusion is much lower resulting in the optimum HETP value for CEC being better by about a factor of two as compared to capillary HPLC (see Fig. 13). Because of no column backpressure, column length for CEC does not need to be reduced as the particle size for the packing particles diminishes. For example, if a 50-cm capillary HPLC column packed with 5-μm particles can provide 45,000 plates, the same CEC column will be expected to give 90,000. However, if the particle diameter is reduced to 1.5 μm, the capillary HPLC column can be only 15 cm in length providing 33,000 plates. The CEC column packed with 1.5-μm particles can still be 50-cm long and now 210,000 plates are available. A representative separation by CEC is shown in Fig. 14.

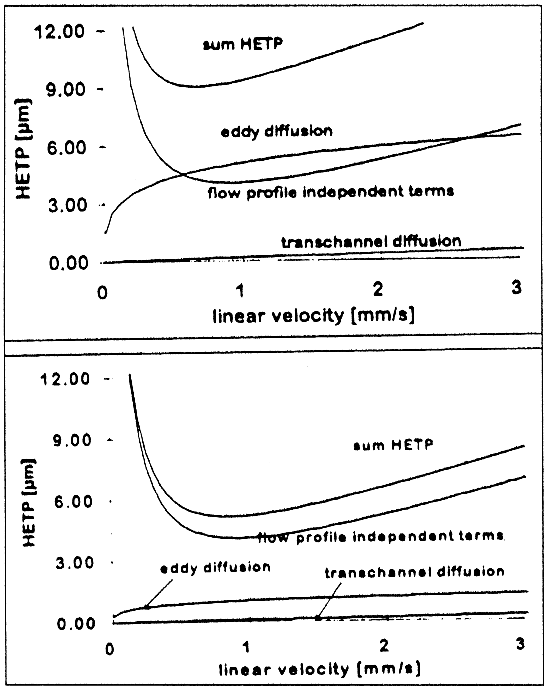


FIGURE 13 Plots of HETP vs linear velocity for HPLC (top) and CEC (bottom) for 5-μm particles. Plots are calculated for an analyte with $k' = 5$ and a diffusion coefficient of 1.5×10^{-5} cm^2/s. [From Dittman, M. M., and Rozing, G. P. J. (1996). *Chromatogr. A* **744,** 63–74.]

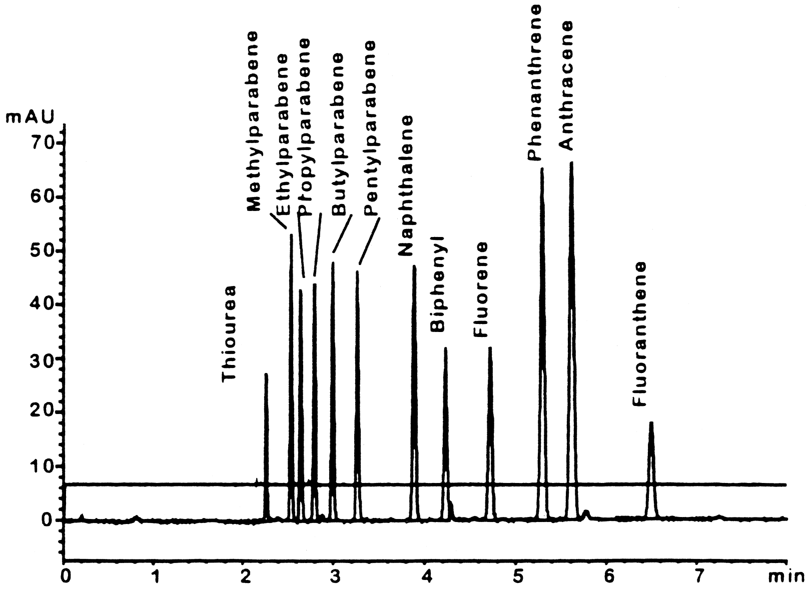


FIGURE 14 Separation of a model mixture containing 5 parabenes, 6 PAHs, and thiourea as the unretained component. Conditions: 250 mm to the detector (335 mm total length) × 0.1 mm CEC Hypersil C-18, 2.5 μm, acetonitrile-25 mM MES, pH = 6 (80 : 20), 20 kV, 10-bar pressure applied to both ends of capillary. Plate numbers 60,000–77,000. [From Dittman, M. M., and Rozing, G. P. J. (1996). *Chromatogr. A* **744,** 63–74.]

C. Preparative HPLC

Preparative HPLC use 25-cm columns with large diameters of 25–150 mm while semipreparative HPLC columns of the same length but 10–30 mm in diameter are used to separate 0.05–1 g samples. Generally the column packing is the more inexpensive 40–50-μm silica and flow rates anywhere from 20 to 300 ml/min are considered reasonable. Because gram quantities of the sample are injected, the detector need not be very sensitive and therefore the RI detector is commonly employed. The goal of preparative HPLC, often in support of organic synthesis research, is to separate as much sample as possible into its components. Therefore, peak resolution is often sacrificed to permit larger injected samples, even to the point of column overload. The chromatography triangle of speed, resolution, and sample capacity indicates that if capacity is of great interest, speed will also likely be sacrificed. As long as the peaks can be distinguished for fraction collection, the separation is considered adequate. Often analytical HPLC, NMR, and IR are used to confirm the purity of the collected fractions. An example chromatogram comparing analytical and preparative HPLC is shown in Fig. 15.

III. SEPARATION TECHNIQUES

The different types of liquid chromatography can be classified into four main classes based on the solute-stationary phase interaction. These are (1) adsorption, (2) partition, (3) ion exchange, and (4) size exclusion. For the first three, the liquid mobile phase has a major role in governing solute retention. The fundamental principles and important packings for each of these LC modes will be described as well as providing typical application chromatograms. In addition, chiral separations will be discussed in a separate section.

A. Adsorption LC

Adsorption LC or liquid–solid chromatography (LSC) is principally carried out in the normal phase mode. It involves no partitioning of the sample solute in the stationary phase. Instead, the polar groups of each organic solute interact through primarily hydrogen bonding forces with the polar sites of the stationary phase. Therefore, careful adjustment of the polarity of the mobile phase for stable activity of the polar sites is needed for reproducible separation.

The most common packing materials for LSC are porous silica $(SiO_2)_x$ or alumina $(Al_2O_3)_x$. Both of these materials have numerous surface hydroxyls that act as the adsorption sites. Silica particles have a high surface area averaging 400 m^2/g and are quite acidic in nature with a surface pH of about 5. This pH is usually not so low as to cause acid-catalyzed solute degradation reactions during the chromatography. As expected, retention of organic bases such as various anilines or nitrogen heterocycles is particularly good on silica. Alumina on the other hand, is

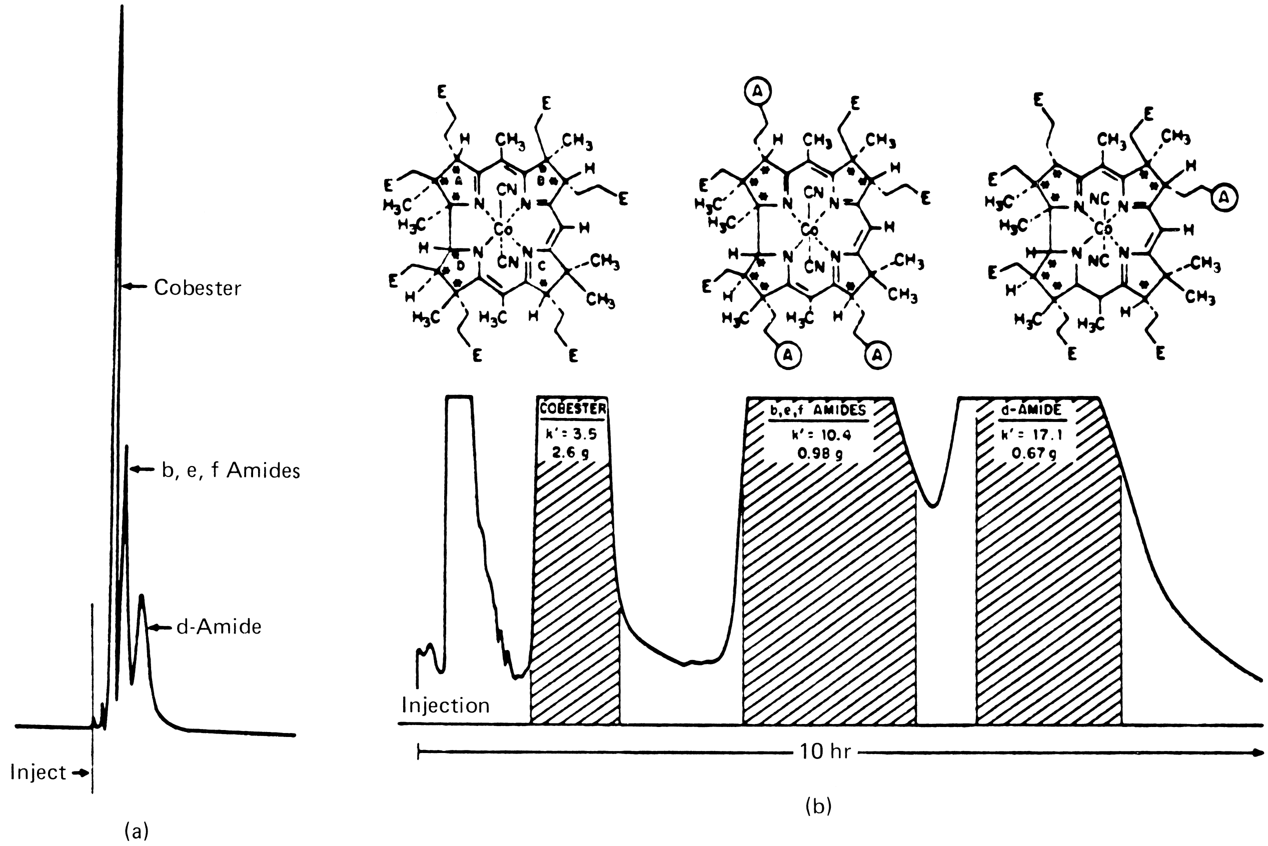


FIGURE 15 Analytical (a) and preparative (b) isolation of vitamin B-12 intermediates. (a) Column, 180 × 0.2-cm i.d., Corasil II, 37–50 μm; mobile phase, hexane/isopropanol/methanol. (b) Column, 240 × 2.3 cm i.d. 37–80 μm silica: mobile phase, hexane/isopropanol/methanol (5 : 2 : 1), flow rate 34 ml/min; injected sample, 5 g. [From Snyder, L. R., and Kirkland, J. J. (1979). "Introduction to Modern Liquid Chromatography, 2nd ed." Wiley, New York, p. 655. Reprinted with permission.]

quite basic with a pH of about 12 and generally is lower in surface area with larger pores. Base catalyzed degradation reactions on alumina can be a problem. However, good retention of acidic organic compounds such as phenols and carboxylic acids is possible on alumina. Florisil, a magnesia–silica coprecipitate, which is strongly acidic in nature, has also been used for LSC. However, silica accounts for about 80% of all applications. Free (non-hydrogen bonded) hydroxyl groups are more reactive to solute polar groups and provide most of the retention. The siloxane, Si–O–Si, bonds are very weak in their adsorption properties. The presence of a polar solvent in the mobile phase such as water will promote hydrogen bonding, decreasing the number of active sites available for solute retention. The LSC retention mechanism can be summarized as a competition between the solute molecules (Z) and the solvent molecules (S) for the adsorption sites.

$$Z_m + nS_{ads} \rightleftharpoons Z_{ads} + nS_m,$$

where Z_m represents solute molecules in the mobile phase, S_{ads} solvent molecules adsorbed on the packing, Z_{ads} solute molecules adsorbed on the packing, and S_m solvent molecules in the mobile phase.

A quantitative log relationship between retention factor k' and mobile phase strength N_B follows.

$$\log k' = k'B - (A_x/n_B)\log N_B,$$

where k'_B = retention factor in a pure nonpolar solvent, A_x = adsorption cross section of analyte X, n_B = adsorption cross section of solvent molecule, and N_B = number of polar solvent molecules. If the slope representing the number of analyte molecules/number of solvent molecules displaced is large, B is either a very polar solvent and/or the analyte is weakly retained. The converse is true if the slope is small.

This adsorption–desorption equilibrium is in operation continuously as the solute molecules pass down the column. The more polar the mobile phase, the more adsorption sites will be blocked by the solvent, causing the solute molecules to remain in the mobile phase and decreasing retention.

TABLE III Properties of Solvents for HPLC[a]

Solvent	ε^{ob}	UV cutof (nm)	Viscosity at 25°C [centipoise (cP)]
Pentane	0.0	195	0.22
Isooctane	0.01	197	0.47
Cyclohexane	0.04	200	0.90
Carbon tetrachloride	0.18	265	0.90
p-Xylene	0.26	290	0.60
Toluene	0.29	285	0.55
Benzene	0.32	280	0.60
Ethyl ether	0.38	218	0.24
Chloroform	0.40	245	0.53
Methylene chloride	0.42	233	0.41
Tetrahydrofuran	0.45	212	0.46
Acetone	0.56	330	0.30
Ethyl acetate	0.58	256	0.43
Aniline	0.62	310	3.8
Acetonitrile	0.65	190	0.34
Dimethylsulfoxide	0.75	268	2.0
Isopropanol	0.82	205	1.9
Ethanol	0.88	210	1.1
Methanol	0.95	205	0.54
Water	large	191	0.90

[a] Most of this data was taken with permission from a similar table in Snyder, L. R., and Kirkland, J. J. (1979). "Introduction to Modern Liquid Chromatography," Wiley, New York, p. 248.

[b] Eluotropic series for alumina (similar rank for silica).

A classification of solvents to their ability to adsorb on the stationary phase is called an eluotropic series (Table III). Retention of solute is reduced with mobile phase solvents of higher solvent strength parameter, ε^0. Solvent polarity parameters (ϕ) are similar to ε^0 values. They can be used to estimate the overall polarity of a binary solvent as $P'_{AB} = \phi_A P'_A + \phi_B P'_B$. For example, ϕ values for hexane, diethyl ether, tetrahydrofuran, ethyl acetate, acetonitrile, and water are respectively 0.1, 2.8, 4.0, 4.4, 5.8, and 10.2. The solvent polarity P'_2 required for a desired k'_2 can be predicted from k'_1 and that solvent polarity P'_1 using the equation $k'_2/k'_1 = 10^{(P'1-P'2)/2}$. A two-fold change in P' results in a ten-fold change in k'. One of the problems of adsorption LC is that solvent impurities (particularly water) in organic solvents can markedly affect solute retention and cause nonreproducible chromatograms. To alleviate this problem and also help reduce peak tailing, the mobile phase can be intentionally saturated with water. Alternatively, addition of a polar organic solvent at less than 1% will also work. Generally, alkanes with either chlorinated, ether, or ester solvents as modifiers are used as mobile phases for LSC. Hexane modified with 50% methylene chloride and 0.1% isopropanol or acetonitrile is considered a good mobile phase to start with.

The usual order or elution of organic solutes is dependent on the type of polar functional groups, number of groups, and orientation. A listing of functional groups from low k' (capacity factor which is proportional to retention) to high k' follows: Alkane < olefins < aromatic ≈ organic halides < sulfides < ethers < nitro compounds < esters ≈ aldehydes ≈ ketones < alcohols ≈ amines < sulfones < sulfoxides < amides < carboxylic acids. As expected, this order roughly reflects the eluotropic series in Table I. A greater number of polar groups will promote retention unless their close proximity permits intramolecular hydrogen bonding.

One of the strengths of LSC is its ability to separate isomers, particularly aromatics functionalized with polar groups, in the retention order ortho < meta < para. The ortho compound is retained the least due to intramolecular hydrogen bonding. The meta functional groups can independently interact with the stationary adsorption sites but not often at the same time. The para isomer is retained longest because the two opposite functional groups can "sit down" on the adsorption surface and both simultaneously interact with the active sites. A chromatogram of nitroaniline isomers is shown in Fig. 16. As required for all types of liquid chromatography, the sample must be soluble in the mobile phase. Therefore, LSC is generally used for organic solvent extracts of solid or aqueous samples as well as characterization of product solutions from organic synthesis.

B. Partition LC

Partition LC or liquid–liquid chromatography (LLC) involves solvation of the solute molecules in the stationary phase held by the packing or solid support. The versatility of partition LC is due to the wide variety of possible stationary phases. Partition LC, like adsorption LC, can be used in the normal phase mode but is more commonly employed for reversed-phase LC which uses a nonpolar stationary phase and a polar mobile phase. The retention between solute and stationary phases can be due to hydrogen bonding, dipole–dipole, and/or Van der Waal forces. Hydrogen bonding forces have been previously described for LSC. Dipole–dipole interactions are electrostatic in nature due to the charge asymmetry of the solute and stationary phases. Van der Waal forces, which dominate in reversed-phase HPLC, are interactions between hydrophobic or nonpolar groups of the solute and the liquid phase. Essentially, the water or miscible organic solvent molecules exist in a high-energy state when adsorbed to the nonpolar (C-18) derivatized silica surface. A larger aromatic solute molecule will preferentially displace many adsorbed solvent molecules in an entropy driven process, resulting in a lower energy state. A general rule of thumb is the

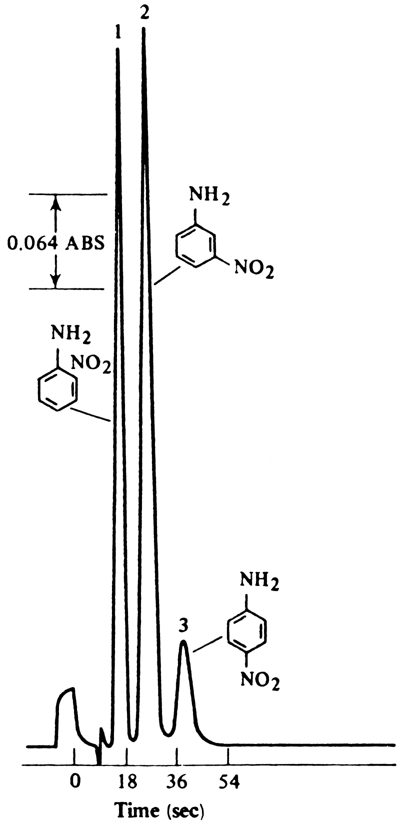


FIGURE 16 LSC separation of nitroaniline isomers on 10-μm alumina, 15 cm × 2.4 mm column, 40% CH_2Cl_2 in hexane mobile phase, flow rate 1.7 ml/min, 1 μg of each isomer. [From Majors, R. E. (1973). *Anal. Chem.* **45,** 757. Reprinted with permission by the American Chemical Society.]

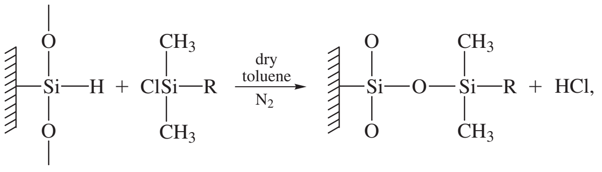
greater the number of hydrophobic groups (CH_3 or CH_2) or the lower the number of hydrophilic groups, the greater the expected retention. It has been shown that the log of k'_B/k'_A for molecules A and B which differ by the CH_2 group in structure is proportional to the surface tension of the mobile phase. However, the partitioning mechanism of retention is also important as indicated by linear plots of log k' versus number of carbons for a homologous series of solutes. As shown by Dill and Dorsey, the density of the stationary phase cannot be too high to permit entry of the solute between the C-18 chains.

A quantitative relationship between log k' and fraction of the nonpolar solvent B (usually water) in the mobile phase (ϕ_B) is given by

$$\log k' = \log k'_w - S\phi_B,$$

where k'_w = retention factor for solute in pure water, and S = solvent strength parameter (a measure of nonpolarity).

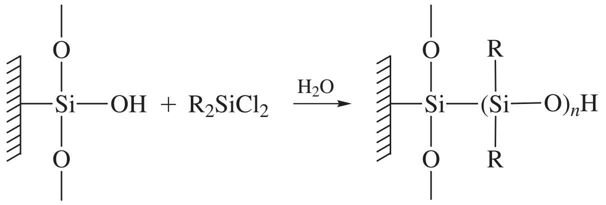
The earliest LLC packings were simply a solid support such as silica coated with the liquid of choice such as oxypropionitrile. The advantages of these packings were ease of preparation, wide choice of liquid phases, and high sample capacity. However, the lifetime and reproducibility of these columns were often poor due to the gradual stripping of the stationary phase by the mobile phase. Saturation of the mobile phase with the stationary phase only partially alleviated the problem. Therefore, bonded-phase packings that have the liquid phase covalently attached to the solid support were developed and now are almost exclusively used. Reversed-phase packings are synthesized by reaction of the desired organochlorosilane with the hydroxyl groups of porous silica to form a siloxane bond as



where R often = $CH_3(CH_2)_3$, $CH_3(CH_2)_7$, $CH_3(CH_2)_{17}$, or phenyl. An organic base such as pyridine is often added to neutralize the HCl produced and drive the reaction to the right. Recently, sonication during the bonding reaction has improved coverage. Ligand loading for a C-18 column is about 2–4 μmol/m^2 silica. Specialty normal phase bonded silica packings such as amino or cyano functionalized materials can be made in an analogous fashion using 4-aminobutyltriethoxysilane and 3-cyanopropyltriethoxysilane. In these reactions, ethanol is produced and the silane reagent can potentially bond to three silica sites. However, if incomplete bonding results, the remaining Si—OCH_2CH_3 moieties will hydrolyze to deleterious Si—OH groups. Unfortunately, reactions of the surface hydroxyls only proceed to an extent of about 50% and the residual Si—OH groups can hydrogen bond with polar groups of solute molecules, causing peak tailing. To partially alleviate this problem, trimethylchlorosilane (TMCS), because of its smaller size, is reacted to "end-cap" many of the remaining hydroxyls. Kirkland and coworkers have found it is important to fully hydroxylate the silica packing before silanization to minimize the number of isolated acidic silanols, which, in particular, cause peak tailing for basic solutes. The C-18 packing is probably the most widely used reversed-phase packing. Gilpin as well as others have shown the orientation of these bonded hydrocarbon chains changes as a function of temperature and solvent. The shorter chain hydrocarbon packings as well as phenyl silica are used when lower retention is desired. Since siloxane bands are cleaved by strong acid, mobile phase pH constraints between 2 and 8 remain a limitation of silica-bonded phase packings.

To improve the lifetime of bonded phase silica packings, reactions with di- or trichloroorganosilanes have been

carried out in the presence of water with the usual resultant formation of a polymeric layer:



These reactions are often difficult to control because both cross-linking and linear polymerization are possible. The polymer layer may be too thick to permit good chromatographic mass transfer or too thin to give adequate sample capacity. In addition, residual silanols will be formed if not all the Si—Cl groups react; an end capping reaction with TMCS is recommended.

Alternatively, bifunctional chlorosilanes with an ether bridging group or simply sterically protected monochlorosilanes such as chlorodiisopropyloctyl silane have both provided protection of the siloxane bond between the silane and silica surface from acid hydrolysis. Using a low pH mobile phase required for the reversed-phase separation of peptides and proteins, essentially no change in column performance was observed between the first and forty-first injection. Polymers have also been cross-linked on the silica surface to form stable packings.

Mobile phases for reversed-phase chromatography are often methanol–water or acetonitrile–water binary mixtures because the organic solvent has good miscibility with water and has a low UV wavelength cut-off. The organic solvent should also have a low viscosity (see Table III) to reduce column backpressure and to minimize the C_m term of the Van Deemter equation by maximizing D_m. This is particularly important because the viscosity of a binary organic solvent–water mixture is generally higher than either the pure solvent or water. Essentially solvent strength as ordered in Table III should be reversed; the greater the polarity of the mobile phase, the stronger the hydrophobic interaction of the nonpolar solute groups with the reversed phase packing. Increasing the water content will enhance retention, while increasing the organic content will reduce retention. Often, a starting mobile phase of 50–50 methanol–water is tried if appropriate mobile phase composition information for a particular sample is lacking. The solvent polarity P'_2 required for a desired k'_2 can be predicted from k'_1, and solvent polarity P'_1 by using the equation $k'_2/k'_1 = 10^{(P'2-P'1)/2}$. Again a two-fold change in P' results in a ten-fold change in k'. An example of reversed-phase HPLC for the separation of beverage additives is shown in Fig. 17.

The water–organic mobile phase should be modified for ionizable solutes. Adding a buffer to control the solvent pH will suppress ionization of either weak organic acids or bases and minimize peak tailing. Strong organic acids and bases often exhibit poor hydrophobic retention and cannot be neutralized in the pH range from 2 to 7.5. For solute anions, a quaternary ammonium salt such as tetrabutylammonium hydrogen sulfate is added to the mobile phase to form an ion pair that can hydrophobically partition with the reversed-phase packing. Ion-pair formation for solute cations is accomplished using a sulfonated alkane such as hexane sulfonic acid. However, the mechanism for ion-pair chromatography is not this simple and immobilization of the ion-pair reagent on the hydrophobic reversed-phase packing with the ionic group oriented out is likely. This *in-situ* ion exchange phase can retain the solute ion through electrostatic means. In any case, ion-pair chromatography is very effective as seen in Fig. 18.

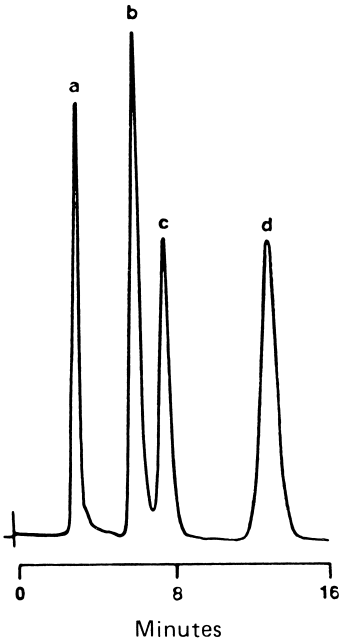


FIGURE 17 Reversed-phase separation of beverage additives on 10-μm C-18 silica (Partisil-10 ODS-2). Column 4.6 mm × 25 cm, mobile phase 50 : 50 methanol–water, flow rate 0.6 ml/min., pressure 529 psi, UV detection at 254 nm. Peaks: (a) Saccharin, (b) Theobromine, (c) Theophylline, and (d) Caffeine. [Reprinted by permission from Whatman, Inc.]

Micellar liquid chromatography is the use of a surfactant such as sodium dodecyl sulfate (SDS) in the mobile phase at a concentration above the critical micelle concentration (CMC) of about 10^{-2} M. At the CMC, aggregation of 60–100 surfactant monomers occurs with the hydrophobic part of the molecule oriented toward the center of the assembly and the hydrophilic tail exposed to the solution. Other surfactants used have been cationic or nonionic in nature, such as cetyltrimethylammonium ion and Brij-35, respectively. For reversed-phase HPLC, the surfactant can

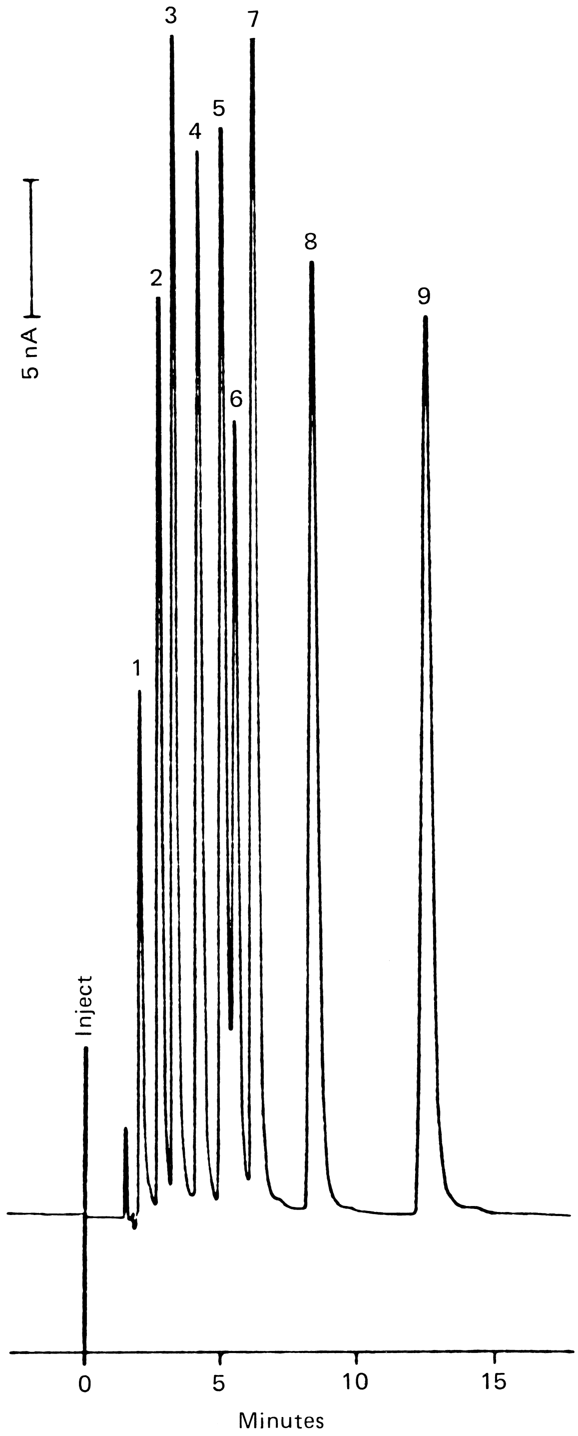


FIGURE 18 Separation of catecholamines and interfering compounds. Ultrasphere C-18 column, 25 cm × 4.6 mm, mobile phase: 10% methanol, 90% 0.1 *M* potassium phosphate, pH 3.0, 0.2 m*M* sodium octylsulfonate, at ±0.72 V vs Ag/AgCl reference electrode, sample size 20 μl. Peaks: (1) Ascorbic acid, (2) Dihydroxyphenylglycol, (3) Norepinephrine (4) Epinephrine, (5) Hydroxymethoxyphenylglycol, (6) Dihydroxybenzylamine, (7) Normetanephrine, (8) Dopamine, and (9) Dihydroxybenzylamine. [Reprinted from permission from Beckman/Altex Scientific.]

replace the typical methanol or acetonitrile modifier. One advantage of micellar liquid chromatography is in gradient elution where the reequilibration time can be dramatically shortened compared to gradient reversed-phase LC. In addition, direct injection of serum samples for drug analysis can be tolerated by the HPLC column if a micellar mobile phase is used.

In general, reversed-phase columns often have only a short lifetime when used for the analysis of drug samples in serum, due to the buildup of proteinaceous material on the hydrophobic particle surface. To overcome this problem, Pinkerton and co-workers developed the internal surface reversed-phase class of packings. This material is synthesized by first bonding a hydrophobic polypeptide containing phenylalanine to the silica surface and inside the pores. Using the enzyme carboxypeptidase A, the phenylalanine groups on the surface are cleaved off, but stearic hindrance prevents any reaction inside the pores. The small drug molecules such as phenobarbital can be still retained chromatographically inside the pores while the protein has little affinity to the surface hydrophilic phase. Recently, modifications of this concept have been directed to attachment of a hydrophilic polymer on the outside surface of the reversed-phase particles to prevent adsorption of proteins. These packings are sometimes called restricted access media (RAM).

Hydrophobic interaction chromatography (HIC) is a type of reversed-phase LC using relatively hydrophilic column packings and a high-salt content in the mobile phase. A dedicated HPLC instrument with titanium pump heads, a special injector, and plastic column hardware with PEEK connecting tubing is recommended to prevent corrosion and delecterious sample interactions caused by stainless steel. Because HIC packings have 10–100 times less density of hydrophobic groups, a high-salt concentration is necessary to enhance the hydrophobic retention. Because proteins retain their native conformation in such mobile phases, HIC is particularly useful for the separation of enzymes without denaturation. Typical experimental conditions are the use of a salt gradient from 2 to 0.1 *M* $(NH_4)_2\ SO_4$, using a propyl or phenyl column for the purification of trypsin with high activity.

Because of the great interest in aqueous samples, a wide variety of reversed-phase HPLC applications have been published. The pharmaceutical, biochemical, food and beverage, and the environmental laboratories represent only a partial listing where reversed-phase HPLC is common.

C. Chiral Separations

Because only one optical isomer of a drug may be pharmacologically active, an important application of HPLC

is in the field of chiral separations. Two approaches can be employed to separate enantiomers. One method is to derivatize the enantiomers with an optically pure chiral reagent, forming two chiral centers in the products. These diasteromers have different physical properties and can be separated by conventional normal-phase HPLC. The derivatizing reagent should have bulky groups attached directly to the chiral center and generate derivatives with the two chiral centers close to each other to provide a more facile resolution of the diasteromers. For example, the reagent a-methyl-p-nitro-benzylamine will permit the resolution of racemic carboxylic acids, while a-naphthylethylisocyanate can modify racemic alcohols before separation.

The second approach is to use either a chiral mobile or stationary phase to directly distinguish the optical isomers. The use of a chiral mobile phase is based on the premise that the sample compounds will form strong associations with the chiral reagent. Based on ligand exchange chromatography, D- and L-amino acids could be separated using an optically active copper (II) proline complex in the mobile phase. If a L-proline ligand is used, the L-amino acid elutes after the D-enantiomer and vice versa using the D-proline ligand. Ion-pair formation using an optically active base such as quinine has permitted the separation of acid enantiomers. In this case, the formation of an optically active dynamic ion exchange resin may also assist in the separation.

For chiral recognition, three simultaneous interactions, one of which is stereochemically based such as hydrogen bonding, dipole–dipole, and/or dipole-induced dipole of the stationary phase with the analytes, should occur. The preparation and characterization of chiral stationary phases for the separation of enantiomers by HPLC has been studied thoroughly by Pirkle and co-workers. For example, chiral N-(3,5-dinitrobenzoyl)-phenylglycine is reacted with aminopropyl silica to form a chiral packing material (Fig. 19). The electron withdrawing dinitrobenzoyl group is a good π electron acceptor favoring the separation of enantiomers with aromatic groups such as N-acetylated α-arylalkylamines. In contrast, the application of a (s)-N-1-N-naphthyl-leucine chiral phase is particularly good to set up an electron-donating type interaction. Separation of dinitrobenzoyl derivatives of amines or thiols is possible. A second type of a chiral stationary phase depends on a size exclusion mechanism. For example, a B-cyclodextrin having a molecular weight of 1000 and 35 chiral centers has secondary hydroxyl groups on the edge of the "donut" structure to preferentially hydrogen bond with an enantiomer of the right configuration Fig. 20. Dansylated D-amino acids such as phenylalanine and leucine have capacity factor values of about four compared to three for the corresponding L-amino acid. The

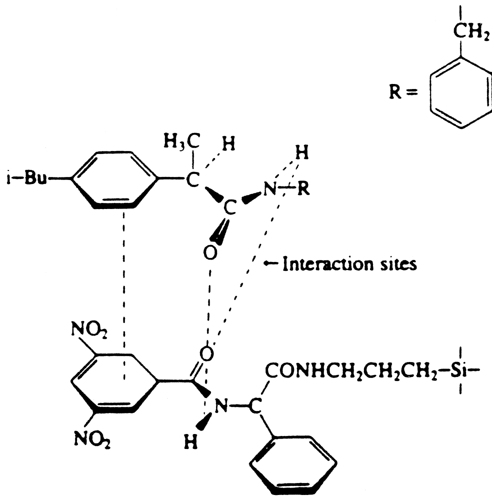


FIGURE 19 Interaction between chiral stationary phase and amide derivative of (R)-ibuprofen. [From Braithwaite, A., and Smith, F. J. (1996). "Chromatographic Methods, 5th Ed." Chapman & Hall, London.]

presence of an aromatic group as part of the solute structure to ensure inclusion complexation with the glycosidic oxygens is important. The other cyclodextrins shown in Fig. 20 either smaller or larger in size can also provide steric chiral recognition but are not as commonly used as the β form. Proteins such as bovine serum albumin (BSA) when bonded to silica have also been shown to provide chiral recognition of low molecular compounds, such as aromatic amino acids, coumarins and benzoin derivatives.

D. Ion-Exchange LC

Ion-exchange chromatography is still considered the dominant HPLC method for the separation of either inorganic or organic ions, particularly the former. The separation mechanism can be best explained as an equilibrium process between the charged functional groups of the stationary phase and the oppositely charged counter ions in the mobile phase as well as the solute ions. The appropriate cation and anion exchange reactions can be written as follows.

Cation exchange:

$$X^+ + \mathrm{Res}^- Y^+ \rightleftharpoons Y^+ + \mathrm{Res}^- X^+$$

Anion exchange:

$$X^- + \mathrm{Res}^+ Y^- \rightleftharpoons Y^- + \mathrm{Res}^+ X^-,$$

where X represents the sample ion, Y the mobile phase ion (counter ion), and Res^+ or Res^- the ionic site on the stationary support resin. A quantitative relationship between $\log k'$ and log of the concentration of the ionic eluent $[E_x]$ is given by

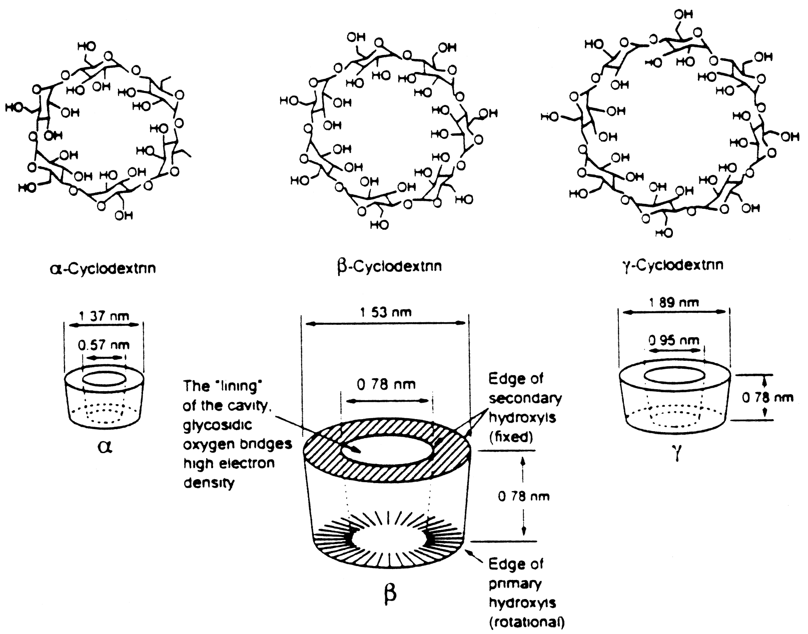


FIGURE 20 Three-dimensional presentation of the geometries of cyclodextrin. [From Braithwaite, A., and Smith, F. J. (1996). "Chromatographic Methods, 5th Ed." Chapman & Hall, London.]

$$\log k' = -(y/x)\log[E_x] + (\log B)/x,$$

where $y =$ charge of the solute, $x =$ charge of the eluent, and $B =$ the product of the capacity of the packing and the equilibrium constant for the ion-exchange process.

Although silica has been used, the most common ion-exchange supports are PS–DVB resins because of their stability at pH extremes. The non-cross-linked benzene rings are available for functionalization. Sulfonation of PS–DVB resin yields the strong cation exchanger, Res–$SO_3^-X^+$, while chloromethylation and subsequent amination forms the strong anion exchanger, Res–CH_2–$N^+(CH_3)_3X^-$. The capacity of these resins, the number of exchangeable groups per gram of resin, can range from 0.1 to 2 meq/g, depending on reaction conditions. Surface agglomerization is a convenient method to prepare low capacity ion exchange packings for ion chromatography. For example, sulfonated PS–DVB microspheres (5–40 μm) are contacted with colloidal anion exchange particles (100–1000 Å) to electrostatically form a surface agglomerated anion exchange resin. The ion-exchange capacity of the resin can be controlled by changing either the size of the microspheres or of the colloidal particles, as well as the degree of functionalization of the latter particles. The corresponding weak anion exchanger Res–$NH^+(CH_3)_2X^-$ and cation exchanger Res–COO^-X^+ have also been developed for use in the separation of labile molecules such as proteins.

The mobile-phase factors of pH and ionic strength primarily control the retention of ion exchange resins. Therefore, buffered solutions are almost always the major component of a mobile phase for ion-exchange LC. For weak acidic or basic solutes, the mobile phase pH controls their ionized state and ability to interact with the resin. The capacity of weak ion-exchange resins is in addition influenced by pH. All other factors considered equal, the greater the capacity of the resin, the greater the ion retention. Finally, the pH as well as the buffer salt can contribute significantly to the overall ionic strength of the mobile phase. Ionic strength is calculated by taking one-half of the sum of the ion concentration times their charges squared. As the ionic strength increases, the amount of counter ion in the mobile phase increases, driving the equilibrium back to the left. This competition of the counter ions for the stationary ionic sites results in a reduced retention of the solute ions. The lower the resin capacity, the smaller the ionic strength that is required to elute the solute ions from the column. The ionic strength is often intentionally increased gradually to improve the separation of weakly and strongly retained ions in a mixture (see Section IV, Fig. 26).

The nature of the ionic solutes often affects their ion exchange retention. As expected, polyvalent ions are held more tightly than monovalent ions. Within a given charge group, retention generally increases with the size of the ion but decreases with the size of the hydrated radius. Solvated ionic radii limit coulometric interactions between ions and energy must be put into the system to strip the water away. The retention order for the alkali metals is $Cs^+ > Rb^+ > K^+ > NH_4^+ > Na^+ > H^+ > Li^+$. Because of its greater hydration, Li^+ is retained less than H^+.

In past years, the direct detection of inorganic anions, cations, and small aliphatic organic acids and bases after column separation has been difficult. Development of ion chromatography in the mid 1970s solved this problem. Now two approaches, both using low-capacity ion-exchange columns and a conductivity detector, are commercially available. First developed by Small, the dual-column ion chromatography system traps the ions of the mobile phase by connecting a high-capacity suppressor column downstream from the analytical ion-exchange column. For example, using a sodium hydroxide mobile phase, the separated anions elute into a protonated cation suppressor column. There the mobile phase is neutralized to water as shown in the equation $\text{Res–SO}_3^-\text{H}^+ + \text{Na}^+ + \text{OH}^- \rightarrow \text{Res–SO}_3^-\text{Na}^+ + \text{H}_2\text{O}$, and the separated anions are changed to the corresponding acids, $\text{Res–SO}_3^-\text{H}^+ + \text{M}^+ + \text{A}^- \rightarrow \text{Res–SO}_3^-\text{M}^+ + \text{H}^+ + \text{A}^-$. Sensitive conductivity detection of the separated ions is now possible at the sub-ppm level. The analogous system for cation analysis, in which HCl is the eluent and the suppressor column is an anion exchanger in the hydroxide form, is equally effective. Hollow fibers, and more recently membranes, have been used in place of the suppressor column. A continuous bathing stream of either acid or base eliminates the problem of periodic regeneration of the suppressor column. Now, even this process has been simplified and just water surrounding the membrane is simply allowed to undergo electrolysis to generate the necessary H^+ or OH^- ions. The second method, single-column or nonsuppressed ion chromatography, uses a low capacity (about 0.1 meq/g or less) ion-exchange separation column permitting low ionic strength mobile phases. The conductivity of the mobile phases is electronically zeroed out, permitting detection of only the sample ions. Although the detection limits are not quite as low as the suppressed ion chromatography method, the single-column method can be easily adapted to existing HPLC hardware and is easier to maintain. Applications of ion chromatography for waste water, boiler water, drinking water, and plating bath samples, as well as others are documented in the literature. Two examples of nonsuppressed ion chromatography are shown in Fig. 21.

The indirect detection method for IEC depends on the use of an ionic mobile phase that not only controls the retention of the sample ions but also responds to the detector of choice. For example, consider indirect photometric detection. After the ion-exchange separation and during the elution process, light-absorbing ions in the mobile phase replaced by photometrically inactive injected sample ions will cause a decreased absorbance at the detector and negative peaks to be recorded. For IPC, salicylate or naphthalenedisulfonate has been used for anion separations and Ce(III) or an aromatic amine have been used for IEC of cations. If a fluorescent or an electrochemically active ionic mobile phase such as Ce(III) is used for IEC, indirect fluorometric or electrochemical detection would be possible in an analogous fashion. Indirect detection limits less than 0.1 ppm are fairly comparable to direct conductivity methods.

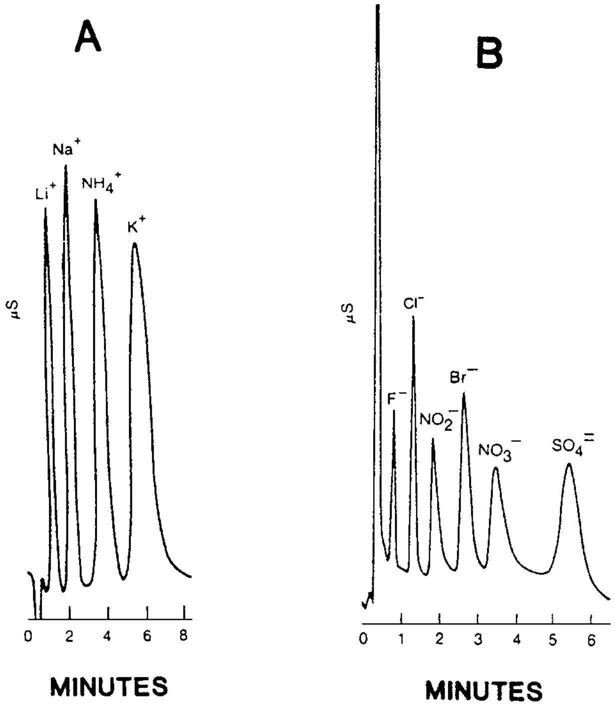


FIGURE 21 Separation of meta ions by ion chromatography. (A) Alkali metals using Ion-200 cation exchange column; eluent: 2.0 mM picolinic acid, pH 2.0; flow rate 2.6 ml/min; sample volume 5 μl; 2–6 ppm each ion; conductivity detection. (B) Inorganic anions using Ion-100 anion exchange column; 1.5 mM phthalate, pH 5.0; flow rate, 1.5 ml/min; sample volume 10 μl; 30–80 ppm each ion; conductivity detection. [Reprinted with permission from Interactions Chemicals, Inc.]

Ion-exclusion chromatography uses an ion-exchange column with an appropriate mobile phase to permit the penetration of nonionic substances into the liquid, both inside and between the resin beads. Retention is based on polar interactions of the solute with the resin functional groups and/or nonpolar forces between the solute and the resin backbone. For example, weak organic acids are separated using a cation-exchange resin and an acidic mobile phase to maintain solute neutrality (Fig. 22). Highly ionized simple inorganic anions pass through unretained. Solutes with a more hydrophobic character such as longer chain hydrocarbon or aromatic monofunctional acids are retained well. In addition, organic acids elute in order of increasing pk_a values. Neutral hydrophilic compounds such as sugars can also be separated by ion exclusion chromatography using a Ca^{2+}- or Pb^{2+}-loaded

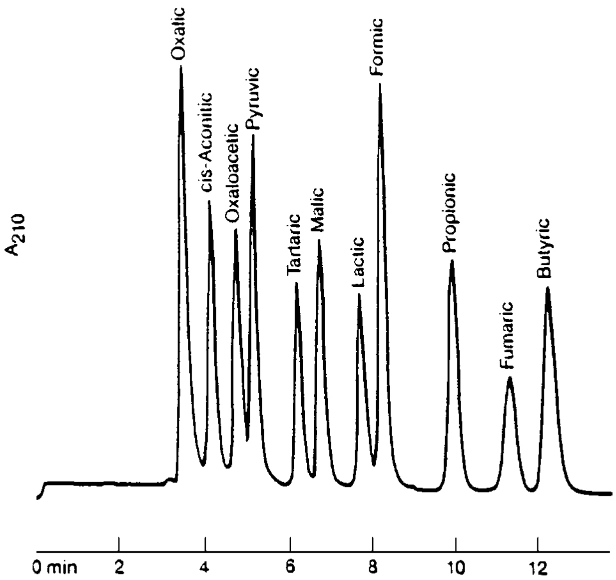


FIGURE 22 Separation of short-chained carboxylic acids by ion-exclusion chromatography. Column: ORH-801 sulfonated cation exchange; eluent: 0.01 N sulfuric acid; flow rate: 0.8 ml/min; temperature: 35°C; detection: UV at 210 nm. [Reprinted with permission from Interactions Chemicals, Inc.]

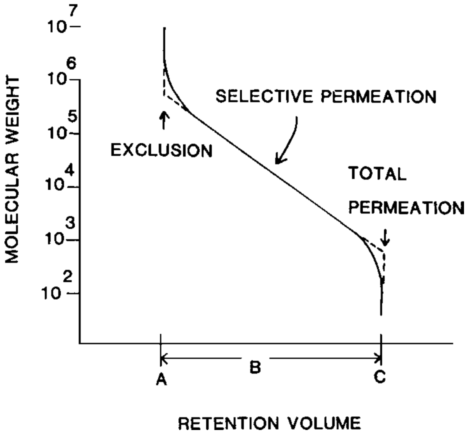


FIGURE 23 Molecular weight calibration curve for SEC.

cation-exchange column and water as the mobile phase. Retention of the sugars is assisted by weak complex formation with the metal cation. Sometimes an inorganic salt is added to the mobile phase to improve retention by promoting a "salting in" phenomenon. Aliphatic alcohols and amines have also been separated by ion-exclusion chromatography.

E. Size-Exclusion LC

Size-exclusion chromatography (SEC) is used for the separation of large-molecular-weight compounds such as polymers or proteins. SEC is generally divided into two classes, gel-filtration chromatography (GFC), which uses aqueous solvents, and gel-permeation chromatography (GPC), which uses organic solvents. The separation mechanism is based on the relative size of the pores of the packing and the molecules to be separated (Fig. 23). If the molecule is large compared to the pore size, it will be excluded from the particles and pass down the column unretained (point A). Molecules similar in size to the pores can partially penetrate the packing particles and are retained to differing extents allowing separation (region B). Molecules much smaller in size than the pores can easily penetrate all the pores of the packing particles and will be retained to the same degree (point C). Therefore, the peaks of an SEC chromatogram are ordered from highest to lowest molecular size. A quantitative relationship for Fig. 19 is $V_r = V_m + K V_s$, where V_m is the volume between the particles, V_s is the volume within the pores of the packing, and K is the partition coefficient described by the ratio of pore volume accessible by the solute divided by the total pore volume. Since separation is based on molecular dimensions (size and shape), monodisperse samples having the same molecular weight (MW) may not be of the same size. Biopolymers such as proteins can adopt different conformations, and small molecules can be associated together depending on the solvent conditions. Therefore, careful column calibration with standards of similar structure is important to obtain reliable MW information of monodisperse samples. For a polydisperse sample as shown by a broad SEC peak, there is no well-defined MW value but instead a distribution of MW values around an average. A number average $\bar{M}n$ or weight average MW, $\bar{M}w$, can be calculated knowing the number and MWs of various fractions of the broad peak. The viscosity detector developed by Yau and the laser light scattering detector have both been shown to be invaluable for obtaining reliable MW information for SEC.

Proper choice of the mobile phase and packing is important to attain a strictly steric retention mechanism. As with other types of LC, both polymers and silica packings have been used for SEC. By controlling the cross-linking during the synthesis of PS–DVB resins, polymers with different pore sizes can be prepared. Because of its hydrophobicity, PS–DVB in normally used for GPC. Sulfonated PS–DVB as well as polyacrylamide are hydrophilic enough to be used for GPC of polar solutes such as sugars. Spherical silica with pore sizes ranging from 60 to 4000 Å are available for separation of molecules from about 500 to 10^5 in molecular weight. Although untreated silica can be used for many sample applications, particularly organic polymers, it is usually modified for the separation of biological molecules. For example, glycophase

silica, Si–O–Si$(CH_2)_3OCH_2$–CH–(OH)CH_2(OH), is preferred for the separation of proteins. The selection of an SEC packing is dependent on the range of components with different molecular weights desired to be separated. Calibration curves of polystyrene standards versus retention volume for packings of different pore sizes are well documented. For 60–Å silica, a linear MW range of 10^2 to 10^4 is possible. For a 750-Å silica, a linear MW range of 10^4 to 10^6 is found. To expand the range of molecular weights that can be separated, it is necessary to connect in series columns of two different pore-sized packings. For the previous example, a linear fractionation range from 10^2 to 10^6 MW would be possible.

Unlike all other modes of LC separation, the mobile phase is not chosen to control peak separation but to ensure sample solubility and minimize solute-stationary phase adsorption effects. Therefore, the sample solvent determines whether the mobile phase is predominantly organic or aqueous. In general, to minimize adsorption effects, a mobile phase that is more strongly adsorbed to the packing than the solute is advised. For example, for the separation of polyurethanes on silica, dimethylformamide is preferred over tetrahydrofuran. The ionic strength should generally be greater than 0.05 *M* when using aqueous mobile phases with silica. One important application for SEC is for the initial exploratory separation of an unknown sample to indicate how complicated it might be. For example, SEC could easily separate a biological sample into high-molecular-weight proteins and low-molecular-weight peptides and amino acids. These peaks could be collected for further HPLC study using ion exchange or reversed phase. A typical example of an SEC protein separation is shown in Fig. 24. Another major use of SEC is for the separation of polymeric oligomers such as polystyrene. In general, the peaks of an SEC chromatogram are quite broad and resolution is only modest. However, the important role that SEC can play in an overall HPLC separation scheme will be elaborated in the next section.

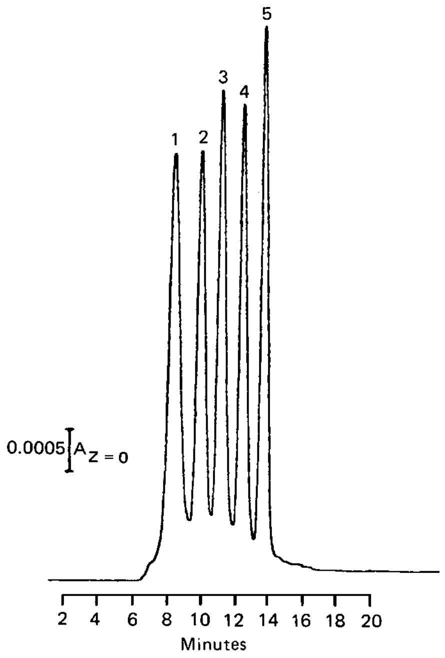


FIGURE 24 Separation of protein standards by SEC. Column is Micropak TSK 3000 SW, 7.5 × 300 mm. Mobile phase is 0.1 *M* potassium phosphate and 0.1 *M* potassium chloride, pH 6.8. Flow rate = 0.7 ml/min. (1) Glutamate dehydrogenase (290,000); (2) lactate dehydrogenase (140,000); (3) enolase (67,000); (4) adenylate kinase (32,000); (5) cytochrome c (12,400). [From Hearn, T. W., Regnier, F. E., and Wehr, C. T. (1983). *American Clinical Products Review*, May/June. Reprinted with permission by International Scientific Communications, Inc.]

IV. DEVELOPMENT OF SEPARATION STRATEGY

A. Choosing a Separation Technique

Knowledge of the physical and chemical properties of a sample is imperative for selecting the preferred separation method or any analytical method for that matter. HPLC can be considered complementary, not competitive, to GC. If the components of a very complex sample can be volatilized (i.e., gasoline), use of capillary GC with MS detection would often be the method of choice. HPLC is better suited to the separation of nonvolatile polar constituents in less complex aqueous samples. Knowledge or experimental determination of the molecular weight, solubility, and ionic nature of the sample is necessary before the appropriate LC method can be selected. Such, a guide to the selection of an HPLC mode is given in Fig. 25. Most separations of a modest number of components can be accomplished using a mobile phase of constant composition. Isocratic elution as compared to a gradient permits better reproducibility and a higher sample throughput as well as the use of simpler equipment. Two primary practical considerations when choosing solvents for a mobile phase are reactivity and purity. The mobile phase should not react with itself or with the sample. Halide salts as mobile-phase additives should be avoided since they can corrode stainless steel tubing and fittings. Impure solvents due to their absorbance at sensitive detector settings can prevent quantitation of trace solutes. Since large volumes of solvent are pumped through the column, trace impurities can also concentrate on the column, impairing its performance. Spectrograde or HPLC grade solvents are highly recommended. In addition, the UV cutoff and viscosity of solvents should be considered (see Table III).

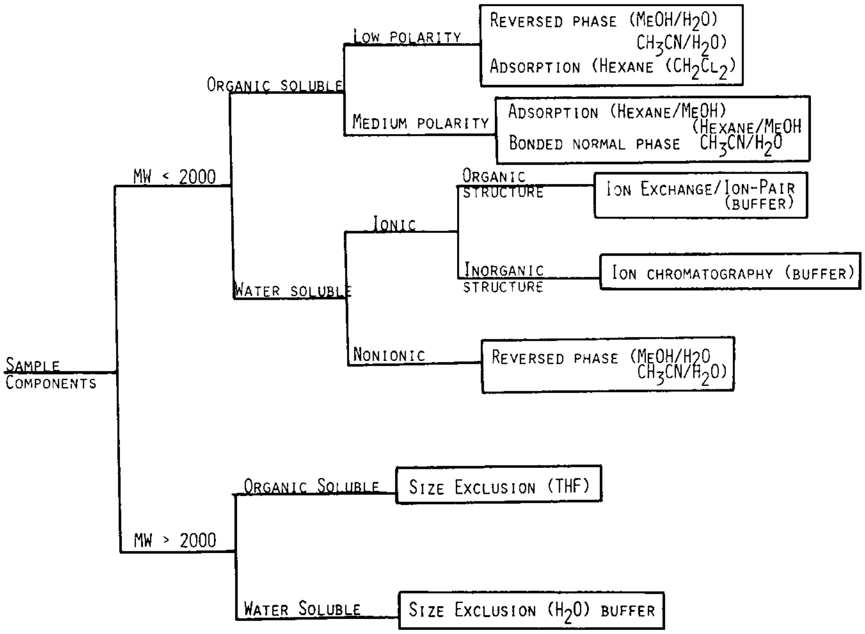


FIGURE 25 Guide to selecting HPLC conditions.

The mobile phase viscosity should be less than 1.0 cP (see Table III) to minimize high column back-pressures. Optimization of the sample separation with respect to resolution, speed, and capacity should be the goal. However, compromise of each of these factors to some degree is necessary; if not, undue cost or labor is expended. Finally, the conditions of an HPLC method can often be decided simply by checking the scientific literature for similar previous studies.

B. Programming an LC Variable

If the sample components have a widely varying degree of retention, separation times can be excessively long for reasonable peak resolution. To improve the separation, programming of some variable during the course of the separation is necessary. The most common parameters are flow rate, solvent strength, temperature, and column type.

Flow programming involves initially using a low flow rate to better resolve the early peaks and then increasing the flow rate to elute well-retained components. It can be carried out step-wise or continuously if desired. Basically, improvement of front-end resolution at the expense of back-end resolution is the result. Decreasing the flow rate by 75% will yield a resolution improvement of about a factor of 1.4. The primary advantages of flow programming are its ease of implementation and no need for column reequilibration between samples.

Solvent strength programming or gradient elution is the most effective and popular programming method. It involves the gradual increase of the mobile-phase solvent strength with time to increase the speed of peak elution. Usually the percent composition of a component of the mobile phase is changed. However, addition of a new stronger solvent can also be done. Gradient elution should be used instead of an isocratic mobile phase if $\Delta T_g / T_g \geq 0.25$, where ΔT_g represents the difference in retention times between the first and last eluting peaks, and T_g is gradient time. The optimum gradient time can be predicted from the equation $T_g = 20\, V_M\, \Delta\phi / F$, where V_M = column dead volume, $\Delta\phi$ = change in volume fraction of the strong solvent during the gradient, and F = flow rate. An average capacity factor of 5 and the isocratic parameter for small molecules equal to 4 must be assumed. The shape of the gradient program can be linear, concave, convex, or stepwise. In choosing a solvent pair, the only requirements are the solvents be nonreactive and miscible. For adsorption LC, methylene chloride is often added to hexane. For reversed-phase LC, acetonitrile is added to water. For ion-exchange LC, a buffer or salt solution is added. A gradient program can be optimized by following these steps: (1) Run a linear program at about 2%-min of the strong solvent B to determine if and where all the components elute, (2) choose the initial % solvent B composition to be half of the %B where the first peak elutes, (3) optimize the program by starting at the front of the chromatogram, inserting sharper %B rates at the appropriate times, and (4) fine-tune the program adding short %B holds to maintain good peak resolution. Computer-assisted methods

such as DRYLAB are commercially available. An example of how gradient elution programming can improve chromatographic time and resolution is shown in Fig. 26.

Temperature programming has been used to a limited extent to control peak retention times. As in GC, an increase in temperature will reduce retention time. This results in sharper peaks with better sensitivity and often an improvement in column efficiency. Because the mobile phase viscosity is reduced, column backpressure drops. For silica-based columns, dramatic increases in temperature from 25 to 70°C have been used to separate peptides and polyaromatic hydrocarbons. High temperatures of 150–200°C have been used for the separation of aromatic compounds in a few minutes using a short polybutadiene (PBD) modified zirconia column with only superheated water as the eluent (see Fig. 27). For polymeric packings, subtle temperature changes in the 35 to 40°C range can reduce peak retention by several minutes. In fact, polymeric columns are often isothermally temperature controlled to improve peak retention reproducibility.

Finally, stationary phase programming of column switching can be used to generate either longer or shorter retention of certain sample components as desired. Schematic diagrams of these common switching modes are shown in Fig. 28. A series arrangement of columns (Fig. 28a). permits good separation of early and late eluting peaks in a reasonable time. The early eluting components are allowed to pass through both columns to the detector while the late eluting components pass only through the first column. The option of eluting the mobile phase to waste (Fig. 28b) is also possible if undesired sample matrix components need to be stripped off rapidly. A parallel arrangement of columns (Fig. 28c) permits true stationary phase programming in which the sample, depending on type and composition, can be directed to column 2 or 3 for complete separation. Finally, the configuration in Fig. 28d permits recycle chromatography or the process of passing the samples through the column repeatedly to improve resolution. This technique is particularly useful in preparative

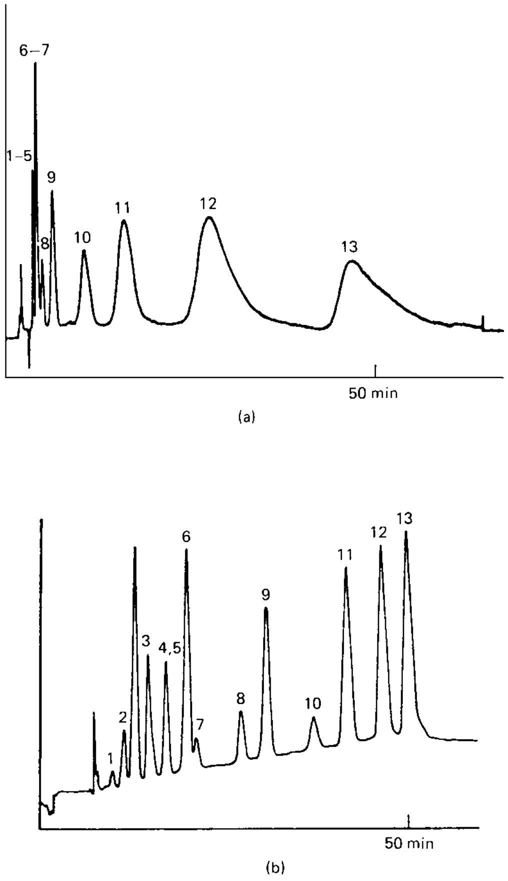


FIGURE 26 Effect of gradient elution on separation. (a) isocratic anion ion exchange, 0.01 *M* borate buffer at pH = 9.7, 0.055 *M* $NaNO_3$. (b) Mobile-phase gradient from 0.01–0.1 *M* $NaNO_3$ at 2%/min. Peaks are (1) o-toluic; (2) benzoic; (3) maleic; (4) phthalic; (5) fumaric; (6) terephthalic; (7) isophthalic; (8) 1,2,3-tricarboxybenzene; (9) 1,2,4-tricarboxybenzene; (10) 1,3,5-tricarboxybenzene; (11) 1,2,4,5-tetracarboxybenzene; (12) pentacarboxybenzene; (13) hexacarboxybenzene. [From Snyder, L. R., and Kirkland, J. J. (1979). "Introduction to Modern Liquid Chromatography, 2nd ed." Wiley, New York, p. 666. Reprinted with permission.]

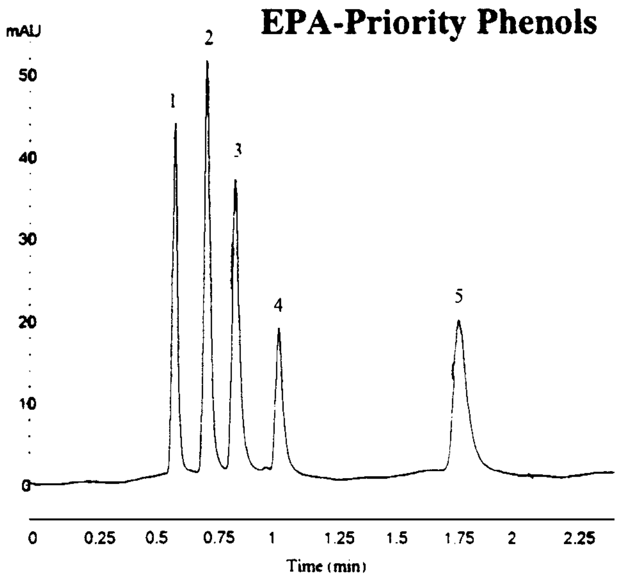


FIGURE 27 High temperature (200°C) separation of phenols using 100% water mobile phase. Conditions: ZirChrom-PBD (150 mm × 4.6 mm i.d.) column, flow rate = 3.0 ml/min, UV detection at 254 nm. Solutes: (1) phenol, (2) 4-chlorophenol, (3) 4-chloro-3-methylphenol, (4) 2,4,6-trimethylphenol, (5) 2,4,6-trichlorophenol. [Reprinted permission from ZirChrom Separations, Inc.]

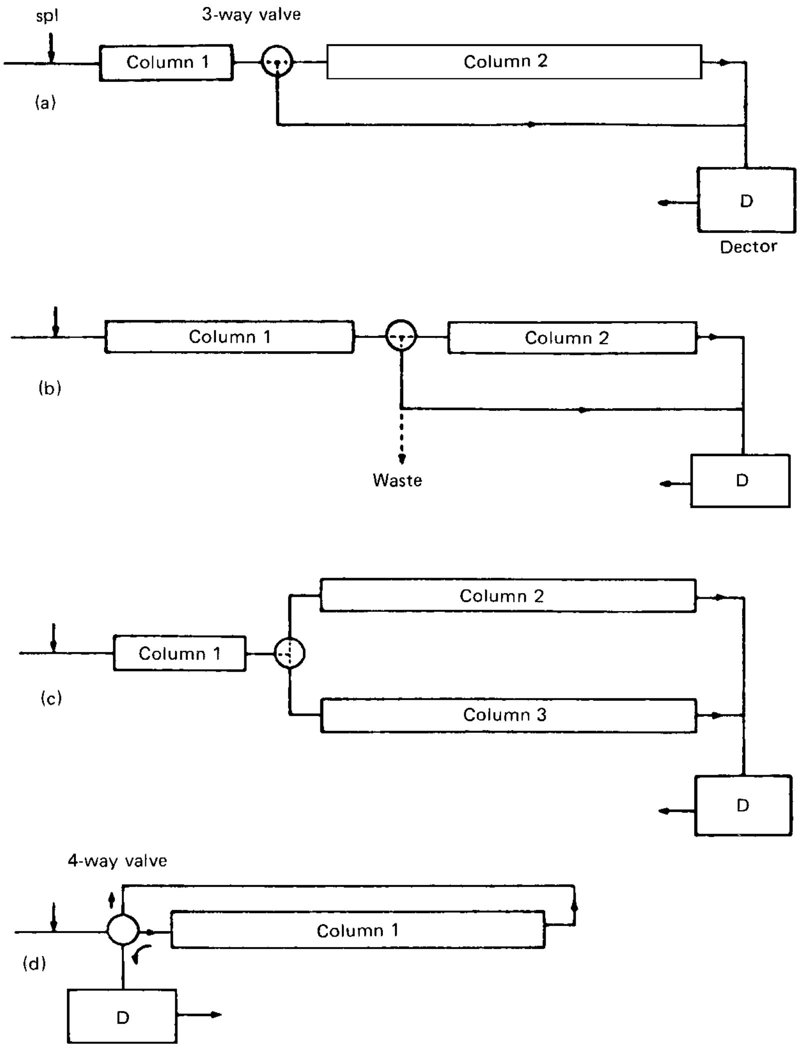


FIGURE 28 Different experimental configurations for column switching. [From Snyder, L. R., and Kirkland, J. J. (1979). "Introduction to Modern Liquid Chromatography, 2nd ed." Wiley, New York, p. 697. Reprinted with permission.]

chromatography. Alternatively, this setup could be used for column backflushing, which involves running the mobile phase backwards through the column. Sample components retained strongly at the front of the column can be eluted easily to the detector. This same procedure can be used to clean the front end of the column of irreversibly retained compounds.

V. SAMPLE DERIVATIZATION CHEMISTRY

The previously described separation strategies are all directed to improving analysis time and peak resolution, assuming the separated components can be detected. However, many compounds of interest, such as aliphatic carboxylic acids or amino acids, are difficult to detect by optical absorbance, fluorescence, or electrochemical means. Precolumn or postcolumn derivatization chemistry is often an inexpensive and effective procedure to remedy this situation. Precolumn derivatization chemistry is carried out prior to and usually separate from the chromatographic operation. The following conditions are desirable for precolumn derivatization.

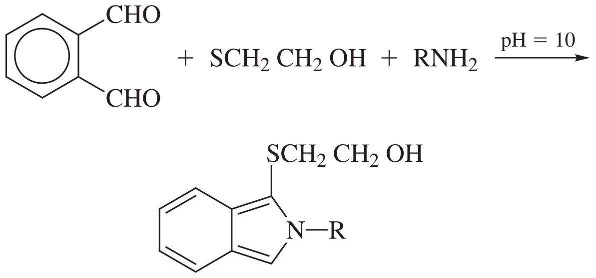
1. Reaction stoichiometry and product structure are known.
2. The reaction should be reasonably fast and the derivatives stable in solution.
3. The derivatives and excess reagent must be readily separable and stable during the chromatographic step.

The primary advantages of precolumn derivatization is that a wide variety of reagents are available for use since

reaction conditions are flexible. Another advantage is that small poorly retained molecules can be modified to more hydrophobic, longer retained compounds. The disadvantage is that sample preparation time is increased and automation is hampered. Because postcolumn derivatization involves mixing the column effluent with the reagent to form the derivative just before it passes into the detector, this method is compatible with automation. Generally, a low-pressure reagent pump, a connecting tee, and a mixing coil are all that are required beyond the basic HPLC instrument. One further advantage of postcolumn derivatization is that the formation of a stable or chromatographically separable product is not necessary. However, fairly stringent conditions are necessary for postcolumn derivatization. These are

1. Reaction must be rapid, within about 2 minutes, and reproducible.
2. The reagent volume should be small relative to the eluent volume to minimize band broadening.
3. The reaction should be a simple one step procedure to minimize additional hardware needs.
4. The reagent itself must have none or a very low detector response.

Most of the HPLC derivatization chemistry is based on well-known reactions reported in organic or inorganic chemistry research. For example, the fluorescent derivatization of primary amines such as amino acids using *o*-phthalaldehyde (OPA) and mercaptoethanol either in the precolumn or postcolumn mode has been well studied.



Scheme 1

A typical chromatogram of OPA amino acid derivatives at the nanogram level is shown in Fig. 29. Dansylation reactions to form fluorescent derivatives of amines or carbonyl compounds have also been applied to HPLC. Complexation of transition metals with pyridylazoresorcinol (PAR) either in the precolumn or postcolumn mode has been a popular method. Precolumn derivatization has been shown by Ian Blair and co-workers to be useful even for LC/MS. Atmospheric pressure chemical ionization commonly used in LC/MS can provide a source of electrons

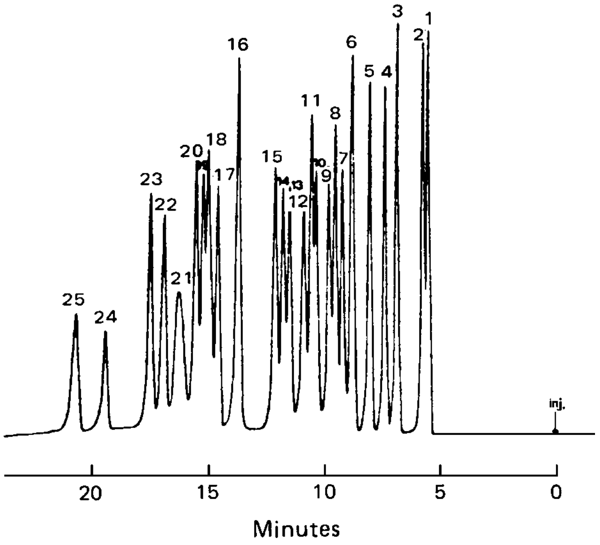


FIGURE 29 Separation of OPA amino acid derivatives on a 5-μm 25-cm C-18 column. Mobile phase: citrate/phosphate buffer, pH 7.7 with convex methanol gradient from 20 to 70% methanol in 15 min. Fluorometer excitation $\lambda = 330$ nm; emission $\lambda = 418$ nm cutoff filter. Peaks: (1) cysteic acid, (2) Asp, (3) Glu, (4) S-carboxymethyl cystein, (5) Asn, (6) Ser, (7) Glu, (8) His, (9) methionine sulfone, (10) Thr, (11) Gly, (12) Arg, (13) β-Ala, (14) Tyr, (15) Ala, (16) α-aminobutyric acid, (17) Trp, (18) Met, (19) Val, (20) Phe, (21) NH_4^+, (22) IDE, (23) Leu, (24) Orn, (25) Lys. [From Lindroth, P., and Mopper, K. (1979). *Anal. Chem.*, **51,** 1668. Reprinted with permission by the American Chemical Society.]

from the nitrogen sheath gas. Steroids, amino acids, or DNA as pentafluorobenzyl derivatives can act as electron capturing compounds to generate negative ions. Detection limits are improved 100 times to attomole levels.

The use of enzymes as precolumn and post-column modification reagents for HPLC has been investigated. For example, nucleoside phosphorylase will catalyze the reaction of orthophosphate and inosine to form hypoxanthine and ribose–6–phosphate. After running the enzyme reaction, separation of the hypoxanthine formed from the substrate inosine by HPLC permits the determination of phosphate in complex matrices. Creatine kinase (CK) isozymes have been separated by HPLC and detected using the luciferase bioluminescence reaction.

$$\text{creatine phosphate} + \text{ADP} \overset{\text{CK}}{\rightleftharpoons} \text{creatine} + \text{ADP}$$

$$\text{ATP} + \text{luciferin} \xrightleftharpoons{\text{luciferase}} \text{adenyl luciferin} + \text{pyrophosphate} + \text{light},$$

where ADP and ATP are adenosine-diphosphate and -triphosphate, respectively.

Postcolumn addition of the substrates of both enzymes and luciferase into a flow cell mounted adjacent to a photomultiplier tube provided sensitive and specific detection of the CK–MB and CK–BB isoenzymes in serum samples.

Chemiluminescence detection continues to receive attention for HPLC postcolumn detection. Molecules that are fluorescent can be detected using a peroxalate reaction scheme. An oxalate derivate, often trichlorophenyloxalate, is oxidized by peroxide or UV light to form a 1,2-dioxetanedione, which can then interact with the fluorescent molecule to generate a CL signal. Detection of dansylamino acids, coumarin modified carboxylic acids, and polyaromatic hydrocarbons in the femtomole range is possible. The metal complex (tris)bipyridine)-ruthenium(III) has also been used for the CL detection of aliphatic amines, particularly tertiary amines, by Danielson and co-workers. Detection limits for drugs such as erythromycin and clindamycin phosphate are in the low ppb range.

Heterogeneous derivatizations, either in the pre- or postcolumn mode, using a reagent as a solid or bonded to packing particles have been explored by Krull and co-workers. Alkyl iodides and epoxides were converted to picryl ethers using silica supported silver picrate before HPLC separation. UV detectability at 220 nm was markedly improved with detection limits at the ppb level. Lower detectability was possible using either reductive or photolysis electrochemical detection.

Postcolumn photochemical derivatization is a relatively simple way to improve the UV or fluorescent properties of a wide variety of aromatic compounds. Irradiation is accomplished on-line by wrapping small diameter Teflon or quartz tubing around a high-energy light source such as a xenon or mercury lamp. For example, cannabinol, only UV active, can be photochemically converted to hydroxylphenanthrene, a highly fluorescent compound. A variety of other drugs such as diethylstilbestrol (DES), clobazam, and tamoxifen can also be irradiated on-line to form fluorescent derivatives. The loss in fluorescent signal as a result of photochemical reaction has been used to confirm the presence of LSD in a chromatogram. A variety of B-lactams, including penicillin and organothiophosphates, have been determined by HPLC with photolysis EC detection.

In the future, undoubtedly, new modifications to the practice of high-performance liquid chromatography, both instrumental and chemical in nature, will be made. The important position HPLC currently holds in the field of separation science shows no sign of weakening and will likely become stronger.

SEE ALSO THE FOLLOWING ARTICLES

Analytical Chemistry • Gas Chromatography • Hydrogen Bond • Organic Chemistry, Compound Detection

BIBLIOGRAPHY

Beesley, T. E., and Scott, R. P. W. (1998). "Chiral Chromatography," Wiley, New York.

Braithwaite, A., and Smith, F. J. (1996). "Chromatographic Methods, 5th Ed," Chapman & Hall, London.

Brown, P. R., and Hartwick, R. A. (1989). "High Performance Liquid Chromatography," Wiley, New York.

Dolan, J. W., and Snyder, L. R. (1989). "Troubleshooting LC Systems," Humana Press, Clifton, New Jersey.

Frei, R. W., and Lawrence, J. F. (eds.) (1982). "Chemical Derivatization in Analytical Chemistry," Vols. 1 and 2, Plenum Press, New York.

Fritz, J. S., Gjerde, D. T., and Pohlandt, C. (1982). "Ion Chromatography," Huthig, Heidelberg.

Horvath, C. (ed.) (1980). "High Performance Liquid Chromatography," Vols. 1, 2, and 3, Academic Press, New York.

Ishi, D. (ed.) (1988). "Introduction to Microscale High Performance Liquid Chromatography," VCH Publishers, New York.

Katz, E. (1998). "Handbook of HPLC," Marcel Dekker, New York.

Kucera, P. (ed.) (1984). "Microcolumn High Performance Liquid Chromatography," Elsevier, Amsterdam.

Lunn, G., and Hellwig, L. C. (1998). "Handbook of Derivatization Reactions for HPLC," Wiley, New York.

Lunn, G., and Schmuff, N. R. (1997). "HPLC Methods for Pharmaceutical Analysis," Wiley, New York.

Mant, C. T., and Hodges, R. S. (1991). "High Performance Liquid Chromatography of Peptides and Proteins: Separation, Analysis, and Conformation," CRC/Lewis Publishers, Boca Raton, Florida.

Neue, U. D. (1997). "HPLC Columns: Theory, Technology, and Practice," Wiley, New York.

Parriott, D. (1993). "A Practical Guide to HPLC Detection," Academic Press, San Diego.

Pasch, H., and Trathnigg, B. (1998). "HPLC of Polymers," Springer, New York.

Rossomando, E. F. (1998). "HPLC in Enzymatic Analysis, 2nd Ed," Wiley, New York.

Sadek, P. C. (2000). "Troubleshooting HPLC Systems: A Benchmark Manual," Wiley, New York.

Snyder, L. R., and Kirkland, J. J. (1979). "Introduction to Modern Liquid Chromatography," Wiley, New York.

Snyder, L. R., Glajch, J. L., and Kirkland, J. J. (1988). "Practical HPLC Method Development," Wiley, New York.

Weiss, J. (1995). "Ion Chromatography, 2nd Ed," VCH Publishers, Weinheim, Germany.

Wu, C. (1995). "Handbook of Size Exclusion Chromatography," Marcel Dekker, New York.

Magnetic Resonance in Medicine

John F. Schenck
GE Corporate Research and Development Center

I. Magnetic Resonance as a Physical Phenomenon
II. Nuclear Magnetic Resonance Scanners as Medical Instruments
III. Imaging Techniques
IV. Clinical Applications of Magnetic Resonance Imaging

GLOSSARY

Bold contrast Blood-oxygen-level-dependent contrast. Contrast mechanism dependent on the level of deoxygenated hemoglobin in the blood. This mechanism is the basis of functional MRI.

Chemical shift Slight shift in the Larmor resonance frequency produced by the magnetic field of the orbiting electrons and varying from one chemical compound to another.

Fourier transform Mathematical process that can be used to analyze a function of time into the individual frequency components that it contains.

Free induction decay (FID) Signal observed from the precessing transverse magnetization after an excitation pulse.

Functional magnetic resonance imaging (fMRI) An imaging technique sensitive to variations in regional blood flow in the brain. This mechanism provides images that demonstrate local brain activation during thought processes and sensory stimulation.

Gradient field Magnetic field that varies linearly in some prescribed direction. Such fields are produced by special gradient coils and are used to create a position-dependent resonance frequency.

Gyromagnetic ratio (γ) Ratio of the magnetic moment to the angular momentum of a nucleus or electron.

Larmor frequency (ω_0) Rate of precession of the nuclear magnetization in an applied magnetic field B; $\omega_0 = \gamma B$ or $f_0 = \gamma B/2\pi$. This is the frequency at which magnetic resonance occurs.

Magnetic field (B) Condition of the space in the vicinity of an electric current or a magnetized material that leads to magnetic forces on moving charges. The SI unit is the tesla (T). An older unit still widely used is the gauss (G) (1 T = 10000 G).

Magnetic moment (m) Measure of the ability of a current loop or a spinning charge to produce a magnetic field.

Magnetization (M) Vector quantity that measures the degree to which a substance is magnetized. It is the total magnetic moment per volume. Longitudinal magnetization is parallel and transverse magnetization is perpendicular to the applied static magnetic field.

Nuclear magnetic resonance (NMR) Technique for detecting nuclear magnetism, utilizes a strong, static field B_0 and a weak, oscillating field B_1 at right angles to it.

Nuclear magnetism Magnetic properties acquired by a substance when there is a net alignment of the nuclear spins.

Pixel Individual picture element. The numerical value of the pixel brightness is used to represent it in the image display.

Precession Motion of a rapidly spinning object when it is subjected to a twisting force or torque. The axis of the spin moves at a constant rate along a conical surface.

Relaxation time Measure of the time required for an excited system to decay or relax toward its equilibrium state. Longitudinal relaxation refers to the longitudinal magnetization and is measured by T_1. Transverse relaxation refers to the transverse magnetization and is measured by T_2.

Resonance Enhanced response of a physical system that occurs when the frequency of an applied driving force is equal or nearly equal to some natural frequency of the system.

Rotating frame Coordinate system rotating around the direction of the static magnetic field at or near the Larmor frequency. It is useful for depicting the motion of the magnetization vector.

Selective excitation Excitation of the spins in a limited region, usually a thin slice, within a sample to be imaged.

Spectrometer Device for resolving, measuring, and characterizing electromagnetic energy produced when matter changes between allowed energy levels. Examples include X-ray spectrometers, optical spectrometers, infrared spectrometers, and NMR spectrometers.

Spin Quantum mechanical property of a nucleus or an electron that gives it an intrinsic angular momentum and magnetic moment.

Spin–warp technique Method of using gradient fields to encode position-dependent information into NMR signals, then using Fourier transform mathematics to decode the signal and produce an image.

Voxel Volume element in the object being imaged. The strength of the signal from a voxel determines the brightness of the corresponding pixel in the image.

NUCLEAR MAGNETIC RESONANCE (NMR) is a subtle physical phenomenon that can be used to observe and exploit a very weak form of magnetism associated with the nuclei of certain chemical elements. Beginning in the late 1970s and early 1980s, it became possible to utilize NMR signals from within human patients as a means of imaging internal disease states for the purpose of medical diagnosis. Although MR scanners were not introduced commercially until about 1982, it is estimated that by 2000 there were more than 10,000 magnetic resonance (MR) scanners in use worldwide and more than 20 million MR scans were being done each year. On the order of 150 million diagnostic MR scans were performed between 1982 and the end of the year 2000.

During the 1980s and 1990s, thousands of clinicians scientists and engineers were engaged in intense efforts to enhance the capabilities of MR scanners by reducing their costs, improving image quality, shortening scan times, and increasing the number of clinically useful applications. These activities have been extremely productive and it seems likely that they will continue for years to come and the number of scanners in service and the number of scans performed each year will continue to increase.

I. MAGNETIC RESONANCE AS A PHYSICAL PHENOMENON

A. Current Status in Medicine

Although the term nuclear magnetic resonance has been in use in the technical community since before World War II, some changes in terminology have accompanied the onset of the medical applications in the 1980s. Partly to avoid any fear on the part of patients that radioactive isotopes might be involved (they are not) and partly to forestall any confusion with the distinct discipline of nuclear medicine (which does utilize radioactive isotopes), the adjective *nuclear* is often dropped. The term NMR is still used in the more technical discussions of the subject, however, as it will be in this article. Two broad general areas of medical applications of NMR have arisen: one is called magnetic resonance imaging (MRI) and the other is called magnetic resonance spectroscopy (MRS). Almost all MRI studies are done using signals from hydrogen nuclei (protons), although in it is possible in certain specialized cases to use signals from other nuclei, such as sodium, to construct MR images. Proton MRI, as the name suggests, utilizes the NMR signal, arising mainly in the protons of the water-containing and fatty tissues of the body, to produce cross-sectional images of the internal anatomy. MRS utilizes the fact that the resonant frequency of a given nucleus varies slightly depending on what chemical molecule it is located within—a phenomenon called the chemical shift. Some MRS studies utilize signals from protons to distinguish chemical compounds, such as lactic acid and the amino acids, that are present in much lower concentrations in the body than is water. But MRS studies more commonly utilize nuclei such as ^{31}P (phosphorus) and ^{13}C (carbon) whose density in tissues is also much lower than that of the water protons. In proton spectroscopy it is

necessary to use special techniques to suppress the very strong water signal that tends to overwhelm the signals from the compounds of interest. Unlike the MRI experiment, the information developed from a MRS experiment generally does not have a sufficient signal-to-noise ratio (SNR) to permit the display of a highly resolved image showing the distribution in the tissue of the nucleus being studied and of the chemical molecules in which it is located. Instead, the data is displayed in the form of a spectrum, which contains peaks associated with various compounds containing the nucleus of interest and originating from a relatively large volume of tissue such as the liver, brain, or muscle.

The distinction between imaging and spectroscopy is more of a convention based on the mode of excitation and display and of the strength of signals detected than it is representative of a fundamental distinction in the physical processes involved. Imaging based on water and fat protons is less technically demanding than most spectroscopy techniques and the results of imaging studies are generally easier to interpret in terms of clinically significant findings. It is equally true, however, that spectroscopy provides more subtle biochemical, rather than anatomical, information on the state of the tissues being studied. It seems fair to say that in 1988 MRI has become a clinically powerful technique already in widespread clinical application while MRS has demonstrated itself as a powerful biomedical research tool with a substantial promise of future clinical application. Both techniques are now the subject of intense research and development activities on a worldwide basis. Because of the more widespread applications at present, this article will focus on MRI.

It should be noted that MRI is but one of several diagnostic imaging modalities that have been developed recently. The others include X-ray computed tomography (the CT scanners), ultrasound, positron emission tomography (the PET scanners), and several nuclear medicine instruments such as single-photon emission computed tomography (SPECT). As a group, these devices have added an enormous new capability to that already provided by conventional X-ray imagers and have revolutionized the medical discipline of diagnostic radiology. Of these instruments the MRI scanners depend on more subtle and generally less-understood physical phenomena than the others; for this reason, this article will consider, in some detail, the physical principles underlying MRI. After a discussion of these physical principles and their historical development, the article will describe the elements of a MR scanner and its safety aspects, the technique by which an image is created from NMR signals, the clinical applications of MRI, and some newer areas of imaging now being developed.

B. Historical Development

It may seem curious that the magnetism of human tissues can be exploited to develop diagnostic information because in normal experience these tissues seem completely unresponsive to magnetic forces. The explanation is that many materials (including water and human tissues), not normally thought of as magnetic, actually possess very weak magnetic properties that are not evident unless special efforts are made to detect them. The magnetic effects involved in medical imaging arise from magnetic properties present in certain atomic nuclei.

The understanding of magnetic properties of materials has developed synergistically along with other basic physical concepts—particularly atomic structure and quantum mechanics—during the twentieth century. In the nineteenth century, chemists developed the concept of an atom as the irreducible, smallest portion of a chemical compound. They established important theoretical concepts such as the periodic table of the chemical elements, and were able to make estimates of atomic size. At this time, however, there was no understanding of the internal structure of the atom, or even a general awareness that such an internal structure existed. Obviously, the concept of nuclear magnetism was not possible before E. Rutherford's experiments, published in 1911, lead to the concept of the nuclear atom illustrated in the familiar cartoon form in Fig. 1. This model conceives of an atom as consisting of

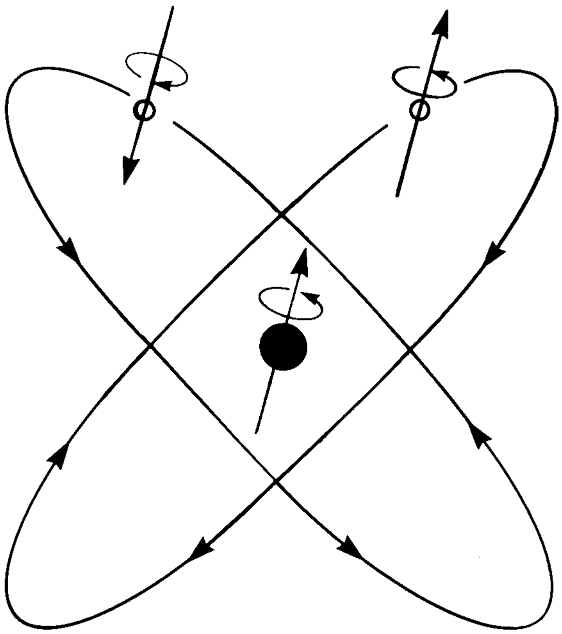


FIGURE 1 Atomic structure. This simple pictorial representation of the structure of a hypothetical atom indicates the three sources of magnetic effects: the orbital motion of the electrons and the spinning motion of the electrons and the nucleus.

a small, dense, positively charged nucleus around which orbit light, negatively charged electrons. The nuclear atom is reminiscent of the solar system, with planets orbiting around the sun. The model has certain defects, which were removed with the development of quantum theory. Unfortunately, the quantum theory replaced the previous concepts of atomic structure with abstruse mathematical concepts, and these cannot be depicted in any fashion that is at once rigorously correct and also visually informative. In other words, quantum theory does not permit pictorial representations of atoms to be taken literally. Nonetheless, the concept of the nuclear atom still provides a convenient and useful aid, particularly in visualizing magnetic phenomena, and we shall make use of it for this reason.

Electric currents (i.e., charges in motion) are known to produce magnetic fields. The nuclear atomic model suggests one obvious source of electrical current—the negatively charged, orbiting electrons—and, this is, in fact, one major source of the magnetic behavior of materials. The other two examples of moving charges within the atom are the electron and nuclear spins.

Much of the experimental information on atomic properties came through detailed study of the light emitted from excited atoms as the orbital electrons changed states, or orbits. Very careful studies of such spectra with spectroscopes of the highest possible resolution showed that the spectra of many atoms were split into several components placed very close together. This phenomenon, called the hyperfine structure, led W. Pauli in 1924 to suggest that atomic nuclei possessed a spinning motion about a central axis. The hyperfine structure was then explained as a weak nuclear magnetic perturbation of the electron energies. In 1925 S. Goudsmit and G. E. Uhlenbeck made a similar proposal, suggesting that the electrons also possess an intrinsic spinning motion, which provides them with both angular momentum and a magnetic moment. Thus the nuclear atomic model postulates three types of electric current that can act as sources of magnetic forces: the orbital motion of the electrons and the spinning motions of the electrons and the nucleus.

The spin magnetic moment of nuclei is usually at least 1000 times smaller than that of electrons, and consequently, the expected strength of the forces produced by nuclear magnetism are extremely small and essentially undetectable. During the 1930s, I. I. Rabi was able to detect and study nuclear magnetic moments of atoms that were moving as atomic beams through an ultrahigh vacuum. An array of magnets was used to slightly deflect the beam by an amount dependent on the nuclear magnetic moment. In 1939, Rabi greatly refined the beam deflection method and reported the first instance of nuclear magnetic resonance. To accomplish this, a weak oscillating magnetic field was added to that of the permanent magnets already in use, and when the frequency of the oscillation matched that of the separation of the nuclear magnetic energy levels a measurable beam deflection occurred. Although the beam method had important applications to fundamental physics, because it permitted the accurate measurement of the strength of the magnetic moments of various nuclei, it was still a very long way from permitting the detection of nuclear magnetism in bulk materials.

An unsuccessful attempt to observe magnetic resonance directly in a bulk material was made in 1936 by the Dutch physicist C. G. Gorter, but he was thwarted by technical difficulties. Many advances in radio-frequency electronics, however, occurred during World War II, particularly as a result of research on radar technology. In 1946, separate groups headed by E. M. Purcell at Harvard and F. Bloch at Stanford detected NMR in bulk materials by use of resonance techniques. This marked the beginning of modern NMR activity. Spectrometers were developed that permitted relatively straightforward NMR experimentation on test-tube-sized samples of a large variety of materials. Because the NMR frequency turns out to be very sensitive to local magnetic fields at the nuclei being studied, it became a major tool in analytic chemistry for molecular structure determination and for the identification of unknown compounds. For example, by 1955 the NMR spectrometers had revolutionized the structural analysis of organic compounds. This application of small-bore spectrometers to analytical chemistry continues unabated to the present time.

At various times several researchers (e.g., J. R. Singer in 1959, T. R. Ligon in 1967, and J. A. Jackson in 1968) reported NMR measurements of one sort or another on human tissues. In a 1971 report, which created substantial interest, R. Damadian described the use of NMR to detect cancer in rat tissues. In the years following that report, a controversy developed regarding the ability of NMR to make a specific distinction between benign and malignant tissues. The tissue is not yet entirely settled, although most workers would agree that, with present capabilities, NMR instruments operating alone cannot make this distinction.

Current work on MRI in humans can be traced to the report of P. C. Lauterbur in 1973 wherein he proposed the use of gradient fields to permit the incorporation of position-dependent information into NMR signals. He created the name zeugmatography to describe the process. Perhaps not surprisingly, the concept has flourished while the name has faded from use (although it is included in the recently published second edition of the *Oxford English Dictionary*). The first NMR images were of small capillary tubes filled with water and placed in an appropriately modified spectrometer. These were soon followed by images of human anatomy. The quality of these images rapidly increased, and many modifications

of Lauterbur's original suggestions were made. In 1977, W. S. Hinshaw and his colleagues published a high-quality wrist image. Their technique used time-dependent gradient fields to localize signal generation to a "sensitive point," which could be scanned through the patient. In 1978, Damadian and associates published an axial cross section of the human chest using a "field focusing" technique that has not been widely utilized in subsequent investigations. By 1980, W. A. Edelstein and co-workers had produced cross-sectional images of the human head and body by the spin-warp technique they had developed. This was a modification of the two-dimensional Fourier transform method proposed in 1975 by A. Kumar and others. The spin-warp technique has become widely standardized in modern NMR scanners. The 1980s witnessed a rapid proliferation of manufacturers and users of MR scanners. These scanners are almost certainly the most significant new diagnostic equipment introduced into medicine during this time.

Thus, in the 70-year period from 1911 to 1981, NMR went through the stages of being an unsuspected property of unknown subatomic structures, through a period of importance only to basic physics, followed by a period during which it was developed as a tool for specialized chemical applications, and finally becoming a crucial medical tool useful to millions of patients and employing thousands of workers.

C. Nuclear Magnetism

Nuclear magnetism is an aggregate property of enormous numbers of identical nuclei responding in step to externally imposed magnetic fields. All magnetic phenomena occurring within matter, in the final analysis, can be explained only by quantum mechanical methods: analyses based on strictly classical methods are known to lead to significantly erroneous conclusions. The most useful method of characterizing phenomena of nuclear magnetism, however, involves the use of macroscopic magnetization vector **M**. The behavior of **M** is governed by the Bloch equations, which are classical in form, but which incorporate, in an empirical fashion, the more fundamental principles of quantum theory. The quantum theory of nuclei in a magnetic field, and in thermal equilibrium with their surroundings, is presented now as a prelude to discussing the Bloch equations.

Each different atomic nucleus is viewed as consisting of protons and neutrons. The properties of the nucleus include its charge, mass, and size, as well as its spin. The spin endows the nucleus with an angular momentum **J** and a magnetic moment **m**. The vectors **m** and **J** are proportional to one another $\mathbf{m} = \gamma \mathbf{J}$, where γ, the gyromagnetic ratio, varies from one nucleus to another and presumably reflects some of the details of the internal structure of the nucleus. The spin of the nucleus is characterized by a spin quantum number I. The quantum theory of angular momentum shows that the value of I is fixed for a given nucleus and can have only integer or half-integer values (i.e., $I = 0, \frac{1}{2}, 1, \frac{3}{2}, \ldots$). The magnitude of the nuclear angular momentum **J** is $\sqrt{I(I+1)}\,\hbar$ ($\hbar = 1.055 \times 10^{-34}$ J sec is Planck's constant divided by 2π). The angular momentum in a given direction can only assume the $2I + 1$ discrete values $-I\hbar$, $-(I-1)\hbar$, up to $I\hbar$. Note that these results display one of the peculiar features of the quantum theory; although the total magnitude of the magnetic moment is $\sqrt{I(I+1)}\gamma\hbar$, the maximum projection in the direction of the applied field, which is the only portion that can be physically observed, is the somewhat smaller value $I\gamma\hbar$.

Protons, electrons, and neutrons have been found to each have $I = \frac{1}{2}$. Nuclei with even numbers of both protons and neutrons all have $I = 0$. The nuclei with $I = 0$ have no angular momentum and no magnetic moment and, therefore, cannot produce nuclear magnetism. This accounts for the large number of chemical elements that cannot be studied by NMR. Odd–odd nuclei all have integral values for I while even–odd and odd–even types have half-integral values for I. Table I is a listing of the spin and magnetic

TABLE I Magnetic Properties of Nuclei of Current or Potential Medical Interest[a]

Nucleus	Resonant frequency at 1 T (MHz)	Nuclear spin, I	Magnetic moment (10^{-26} A m^2)	Natural abundance (%)
^{1}H	42.57	$\frac{1}{2}$	2.44	99.985
^{2}H	6.54	1	0.61	0.015
^{3}H	45.41	$\frac{1}{2}$	2.61	0.0
^{13}C	10.71	$\frac{1}{2}$	0.61	1.10
^{14}N	3.08	1	0.29	99.63
^{15}N	4.31	$\frac{1}{2}$	−0.25	0.37
^{17}O	5.77	$\frac{5}{2}$	−1.13	0.038
^{19}F	40.05	$\frac{1}{2}$	2.30	100.0
^{23}Na	11.26	$\frac{3}{2}$	1.45	100.0
^{25}Mg	2.61	$\frac{5}{2}$	−0.51	10.0
^{31}P	17.23	$\frac{1}{2}$	0.99	100.0
^{33}S	3.27	$\frac{3}{2}$	0.42	0.75
^{35}Cl	4.17	$\frac{3}{2}$	0.54	75.7
^{39}K	1.99	$\frac{3}{2}$	0.26	93.258
^{41}K	1.09	$\frac{3}{2}$	0.14	6.73
^{43}Ca	2.86	$\frac{7}{2}$	−0.75	0.135
^{57}Fe	1.36	$\frac{1}{2}$	0.079	2.2
^{127}I	8.51	$\frac{5}{2}$	1.68	100.0

[a] Adapted from data in Walker, F. W., *et al.* (1984). "Chart of the Nuclides," General Electric, San Jose, CA; Lederer, C. M., and Shirley, V. S. (1978). "Table of Isotopes," 7th ed., Wiley, New York.

moment properties of several nuclei of current or potential medical interest. Note that most of the nuclei in this table have a positive magnetic moment that corresponds to a spinning positive charge. Magnetic moments with a negative sign correspond to a spinning negative charge. If a magnetic field **B** is applied to the sample, the nuclei will be given an energy $-\mathbf{m} \cdot \mathbf{B}$. Thus, the energy of a state depends on the orientation of **m**, and therefore **J**, with respect to the applied field.

The projection of the vector **m** in the direction of **B** must take one of the $2I + 1$ values $-\gamma\hbar I$, $-\gamma\hbar(I - 1)$, up to $\gamma\hbar I$. The nucleus, therefore, has available to it $2I + 1$ states with different energies. The energy of these states will be equally spaced from one another by an amount $\Delta E = \gamma\hbar B$. The state with the lowest energy (the most favored state) has the magnetic moment most nearly parallel to the applied field and vice versa.

Quantum mechanics predicts that if an oscillating magnetic field is applied to the spin system, there will be a resonant exchange of energy between the field and the spins when the quantum energy $\hbar\omega$ in the oscillating field corresponds to the separation between adjacent energy levels. Transitions between nonadjacent energy levels are not allowed. This criterion predicts a resonant interaction when $\hbar\omega_0 = \gamma\hbar B$ or, equivalently, $\omega_0 = \gamma B$. It is important to note that this resonant frequency does not depend on either $\hbar$ or I. This ties in with the fact that the nonquantum, classical analysis, to be discussed later, of magnetic moments in a magnetic field gives the same value for the characteristic frequency.

The states with the magnetic moment in the direction of the field have a lower energy than those with the opposite orientation. Consequently, if the spin system can come to equilibrium with its surroundings at a temperature T, the lower energy states will become more populated than the higher energy states and the substance as a whole will take on a net nuclear magnetization. This represents an aggregate effect of the tendency of all of the individual nuclei to orient themselves parallel to the applied magnetic field. This tendency toward alignment is, of course, opposed by the randomizing effects of the thermal energy present in the material.

The difference between energy levels caused by the applied magnetic field is quite small in comparison to the thermal energy. For a system with only two levels ($I = \frac{1}{2}$), the ratio of the number of spins in the lower energy state n_+ to that in the higher energy state n_- is given by

$$n_+/n_- = e^{\Delta E/kT} = e^{\gamma\hbar B/kT}. \tag{1}$$

Here, k is Boltzmann's constant and T is the absolute temperature. At body temperature 37°C or 310 K, $kT = 4.28 \times 10^{-21}$ J. Even for protons, which have a relatively large magnetic moment, and for a rather intense field of $1.5T$, this formula shows that, for every 10 million protons in a sample, there are only 50 more nuclei in the favored, parallel state than in the higher energy, antiparallel state. Other things being equal, it would improve the sensitivity of MRI if this population difference could be increased. Equation (1) shows that this difference can be increased by increasing the field strength or by lowering the temperature. It is clearly impractical to achieve a significant decrease in patient temperature, however, and there are substantial technical difficulties in achieving body-sized magnets much stronger than those currently in use.

When a material is magnetized, the strength and direction of the effect is given by the vector **M**, called the magnetization, which is defined as the total magnetic moment per unit volume. If a region of volume V contains a large number of individual magnetic moments $\mathbf{m}_i$, then $\mathbf{M} = \sum \mathbf{m}_i / V$, where the sum is over all the sources in the region. The dimensions of **m** are amperes times square meters (A m^2); therefore, the dimensions of **M** are amperes per meter. If, as in the present case, the magnetization is proportional to the applied field, the susceptibility χ (which is dimensionless) is defined by the formula

$$\mathbf{M} = \chi \mathbf{B}/\mu_0, \tag{2}$$

where $\mu_0 = 4\pi \times 10^{-7}$ H/m is a constant called the permeability of free space. For any material the total susceptibility will be the sum of the contributions from each of the relevant sources of magnetic moment; the orbital electron motion, the electron spin, and the nuclear spin. In the present case, of course, we are particularly interested in χ_n, the contribution of the nuclei to the total susceptibility. Statistical analysis of the distribution of the nuclei among the available energy states shows that a nucleus with a spin I, magnetic moment m (A m^2), and a density of ρ (spins per cubic meter) will have a nuclear magnetic susceptibility given by

$$\chi_n = \frac{\mu_0 \rho m^2}{3kT}. \tag{3}$$

Pure water has a density ρ of 55 moles/liter, or equivalently, 6.62×10^{28} protons/m^3. Using the values in Table I, the nuclear magnetic susceptibility for the protons in water is found to be 3.86×10^{-9}. The magnetic behavior of water is particularly relevant to MRI because most of the signal derived from human tissues originates from water molecules. Note that m in Table I and Eq. (3) refers to the total magnitude of the magnetic moment vector. Another convention often used is to refer to the maximum observable component of the magnetic moment vector as the magnetic moment. In the notation used here this component is equal to $m\sqrt{I/(I+1)}$.

D. Comparison to Other Forms of Magnetism

As already mentioned, individual atoms can create magnetic fields by three general mechanisms: the orbital motion of their electrons and the spin motions of their electrons and nuclei. In most cases, for any large collection of atoms, there is a very strong cancellation of magnetic properties so that on a macroscopic scale most materials appear to be nonmagnetic. The most important exceptions are the ferromagnetic materials, of which iron is the most familiar. In this case, internal quantum effects force many of the electron spins to be in alignment with one another, and an intense and easily observable spontaneous magnetization results, even when no external field is applied.

Nonferromagnetic materials, such as human tissues, have no net magnetization until placed in an external magnetic field and even then the resultant magnetization is so weak that it cannot be detected except by use of very sensitive equipment. Alignment processes, which cause **M** to oppose the applied field, give negative contributions to the susceptibility and are called diamagnetic. Processes with the opposite behavior are called paramagnetic. Generally speaking, diamagnetism results from a field-induced alteration of the electron orbits. This produces an induced field that tends to oppose the applied field. Paramagnetism results from the tendency of spinning particles (electrons or nuclei) to align with the applied field. The electron spin paramagnetism tends to be small or absent because in most materials these electron spins tend to cancel in pairs. The orbital motion of the electrons in the H_2O molecule give bulk water a diamagnetic susceptibility of -9.05×10^{-6}. From Eq. (3) the paramagnetic nuclear susceptibility of the protons in water at body temperature was calculated as 3.86×10^{-9}. The total magnetic susceptibility of water is the sum of these two values. Even though water has one of the highest nuclear magnetic susceptibilities due to the large magnetic moment of the proton and the large number of protons in a sample of water, the nuclear paramagnetism of water is completed swamped by the orbital diamagnetism of the electrons even though that itself is a very weak effect.

It is useful to compare the relative strengths of the various forms of magnetism. The spontaneous magnetization of iron is about 1.7×10^6 A/m. In a rather intense applied field of 1 T, the bulk magnetization of water would be -7.0 A/m due to the orbital diamagnetism and 3.1×10^{-3} A/m due to the nuclei. These comparative figures explain why it is not possible to detect nuclear magnetization by direct effects such as the orientation of iron filings or the deflection of a compass needle, even though these effects can be easily demonstrated with ferromagnetic materials. The nuclear magnetization is too weak, by several orders of magnitude, to produce these effects. Furthermore, unlike ferromagnetism, nuclear magnetism is present only when an external field is applied.

E. Bloch Equations

The magnetic energy associated with the spin of an individual nucleus is far too weak to produce detectable effect under any circumstances. In practice, we are always concerned with the signals arising from the additive effects of enormous numbers of nuclei (e.g., 1 mm^3 of water contains 6.62×10^{19} protons). It is desirable, therefore, to turn from the properties of individual nuclei governed by quantum mechanics to the behavior in space and time of a statistically averaged quantity (i.e., **M**), the bulk nuclear magnetization. Strictly speaking we should use a more qualified notation, such as $\mathbf{M}_n$, to indicate that we are now concerned only with the nuclear component of the magnetization. Instead, however, we will follow convention, use **M**, and assume that the restriction is clear from the context. Once the statistical averaging is carried out, the peculiarities associated with the quantum behavior of the individual nuclei are no longer evident. For example, unlike the magnetic moment of a single nucleus, the component of the magnetization in a given direction can take on a continuous range of values without restrictions. It is found that the nuclear magnetization can exhibit an elaborate behavior as a function of time that is driven by two factors. First, there are the externally applied fields that act on all of the nuclei equally and simultaneously to produce a gyroscopic motion of the magnetization vector. Second, there is the effect of all the internal magnetic fields that are derived from the electrons and other nuclei in the vicinity of each individual nucleus to produce relaxation processes. These relaxation processes continually drive the magnetization toward its equilibrium value. The effects of the externally applied fields, considered for a moment to be acting alone, are to provide a torque acting on the magnetization such that

$$\frac{d\mathbf{M}}{dt} = \gamma(\mathbf{M} \times \mathbf{B}). \tag{4}$$

This equation does not yet include the effects of the interaction of the spins with their internal surroundings, the "lattice." However, it contains many of the physical concepts that explain how the nuclear magnetization can be detected despite its weak strength. Equation (4) will describe the evolution of **M** for short times (i.e., for times much shorter than T_1 and T_2 described in the following). In human imaging studies this means that Eq. (4) by itself is enough to describe processes that require a few milliseconds or less to complete. This includes the Larmor precession, which completes each cycle in a fraction of a microsecond. It also includes the B_1 excitation pulses,

which normally last a few milliseconds or less. Whenever Eq. (4) is valid, vector calculus states that the time rate of change of **M** must be perpendicular to **M** because of the properties of the vector cross product. This means that, under the action of the externally applied fields, **M** cannot change in length. Therefore, if by some means, a nuclear magnetization has been created within a specimen at an initial time, the externally applied forces will cause the magnetization to move with time, but only in a way that keeps the length of the magnetization vector constant. The simplest situation is when the external field is simply a constant B directed along the z axis. If at $t = 0$ there is a transverse component of **M**, M_t, pointing along the x axis, the solutions to Eq. (4) are

$$M_x = M_t \cos \omega_0 t \qquad \text{and} \qquad M_y = -M_t \sin \omega_0 t,$$

where $\omega_0 = \gamma B$ and M_z is constant. That is, the component of the magnetization along the z axis remains unchanged as time goes on, while the transverse magnetization rotates at a rate (called the Larmor frequency) about the direction of the applied magnetic field. Thus the total **M** vector at any point moves steadily, at a constant rate, sweeping out a cone whose axis is the direction of **B**. This motion is analogous to the motion of a rapidly spinning gyroscope (such as a top) responding to its own weight. This precession is of basic importance to the detection of nuclear magnetism that, as has been previously mentioned, is much too weak to be detected directly by the magnetic forces it exerts. The precession of the transverse magnetization produces a time-dependent magnetic field and, therefore, by Faraday's law, a time-dependent electric field. This electric field can be detected as a voltage in a coil situated outside the sample. It is of fundamental importance that this induced voltage turns out to be large enough, in many cases, that the precessing nuclear magnetization can be detected electronically. Note that, although different lines of reasoning are used, both the quantum mechanical and the macroscopic approaches lead to a characteristic frequency given by $\omega_0 = \gamma B$.

A concept that is often utilized to describe the motion of the magnetization vector is a coordinate system rotating at or near the Larmor frequency. If the new coordinate system is rotating at exactly the Larmor frequency, then in it, for the example previously given, there is no motion of the magnetization vector. To specify the direction of rotation, we note that nuclei with positive values for the gyromagnetic ratio precess in a clockwise direction when viewed from the positive z direction. The advantage of the rotating coordinate system is that it disentangles the very rapid precession motion, which usually takes place at megahertz rates, and that is caused by the strong static field, from the much slower motions produced by weak, superimposed, oscillating external magnetic fields and from the relaxation processes to be described later. These latter effects normally produce changes in **M** that occur on the millisecond to seconds time scale.

The most common approach to creating the magnetic resonance phenomena is to use a strong field, which we shall designate as B_0, in the z direction and to add to it a weaker oscillating field B_1 that is oriented at right angles to the z axis. If the oscillating field has both x and y components, $\mathbf{B} = B_1(\cos \omega_1 t\hat{\imath} - \sin \omega_1 t\hat{\jmath})$, then in a frame rotating at the frequency ω_1 (moving clockwise when viewed from the positive z direction) it is just a constant B_1 along the rotating x direction. Because such a field contains two components at right angles to one another, it is referred as a quadrature excitation field and is said to be circularly polarized with a clockwise rotation. It is not hard to show that a linearly polarized field with twice the amplitude (e.g., $\mathbf{B} = 2B_1 \cos \omega_1 t\hat{\imath}$) will have the same effect on the spins as the circularly polarized field above. Interestingly a field rotating in the "wrong" direction $B = B_1(\cos \omega_1 t\hat{\imath} + \sin \omega_1 t, \hat{\jmath})$ will have essentially no effect on the spins. In this article we will assume that a quadrature B_1 is used since this simplifies the analysis somewhat.

The exact solution of Eq. (4) in these circumstances is not difficult but the results are more complex than we wish to present here. The essential features of the solution are that the B_1 field has a negligible effect on the motion of **M** unless its frequency ω_1 is close to the Larmor frequency ω_0. More specifically, unless ω_1 is within a frequency range γB_1 of ω_0, the oscillating field will be ineffective. If ω_1 is equal to ω_0, the motion in the rotating coordinate system is very simple. Then the magnetization vector will rotate about the B_1 field, which will be constant in this frame at the rate γB_1. Thus, if the magnetization is along the z axis at time $t = 0$, it will rotate about B_1 and will make an angle $\theta = \gamma B_1 t$ with the z axis after time t. Thus, if the oscillating field operates for a time t equal to $\pi/(2\gamma B_1)$, the magnetization will rotate 90° ($\pi/2$ rad) and will be located in the transverse plane. If it operates for twice this time, the magnetization will be rotated 180° and would be inverted from its initial position.

If a sample is placed in a strong magnetic field B_0, the earlier analysis shows that initially the magnetization is zero, but should increase with time to an equilibrium value $M_0 = \chi_n B_0/\mu_0$. The gyroscopic equations cannot describe this process since Eq. (4) shows that the length of the **M** vector cannot be changed by the external fields. The internal fields provide the answer to this paradox as they permit an exchange of energy between the surroundings, referred to somewhat loosely as the lattice, and the nuclear spin system. Because the internal fields are the result of rapid, essentially chaotic, motion of the atoms of a liquid relative to one another, it is extremely difficult to calculate, from

first principles, what those fields are as a function of time, much less their effect on the magnetization. In 1946 Felix Bloch cut the Gordian knot by lumping the total effect of all these internal processes into two parameters called the relaxation times. The effect of one of the parameters called T_1, the spin–lattice relaxation time, is to govern the rate at which M_z approaches it equilibrium value M_0:

$$dM_z/dt = (M_0 - M_z)/T_1. \tag{5}$$

In equilibrium, there is no transverse magnetization, consequently the internal fields must act to reduce any M_x and M_y that may be present. Bloch proposed quantifying this process by using a second relaxation time T_2:

$$dM_x/dt = -M_x/T_2, \qquad dM_y/dt = -M_y/T_2. \tag{6}$$

The reason that different relaxation times are needed to describe transverse and longitudinal relaxation is that the strong external field biases the response of the spin system so strongly that these two magnetization components respond differently to the weak, internal fields. Analysis of the microscopic mechanisms responsible for relaxation show that T_2 will always be shorter than, or at most equal to, T_1.

Bloch conjectured that the total motion of the magnetization vector can be described as the superimposed effects of the gyroscopic motion (driven by the externally applied fields B_0 and B_1) and the relaxation processes (associated with internally generated magnetic fields). This combination leads to the final form for the Bloch equations in the stationary frame with unit vectors $\hat{\imath}, \hat{\jmath}$, and $\hat{k}$:

$$\frac{d\mathbf{M}}{dt} = -\frac{M_x}{T_2}\hat{\imath} - \frac{M_y}{T_2}\hat{\jmath} + \frac{M_0 - M_z}{T_1}\hat{k} + \gamma \mathbf{M} \times [\mathbf{B}_0 + \mathbf{B}_1(t)]. \tag{7}$$

The Bloch equations give a complete description of the behavior of the magnetization within a body. Qualitatively, they express relatively simple ideas, the transverse magnetization $M_x\hat{\imath} + M_y\hat{\jmath}$ is constantly relaxing toward zero while precessing rapidly at the Larmor frequency γB_0. The longitudinal magnetization M_z is constantly relaxing toward its equilibrium value M_0. If B_1 is not zero it is constantly rotating $\mathbf{M}$ about an axis parallel to B_1 in the rotating frame. In practice, the solution to the Bloch equations may be relatively complicated, particularly if the frequency of $B_1(t)$ is not exactly equal to the Larmor frequency.

The relaxation times T_1 and T_2 provide only an empirical treatment of the effects of the internal magnetic fields. In practice, they must be found by experiment rather than by calculation. Once T_1 and T_2 have been determined, experimental results indicate that the Bloch equations provide a completely satisfactory description of all nuclear magnetic resonance phenomena within human tissues. In more general applications the Bloch equations have been found to provide good predictions for the behavior of liquid or liquidlike samples. In solids, however, the Bloch equations require substantial modification to give accurate results. The protons in biological tissues behave, from a magnetic resonance standpoint, as though they were located in a liquid environment.

F. Relaxation Times

The range of values taken by T_1 and T_2 in human tissues is crucial to determining the practicality of MRI for human tissues. If an unmagnetized sample (e.g., a human patient) is placed in an uniform magnetic field, it is initially unmagnetized. The Bloch equations show that the nuclear magnetization will gradually build up along the z direction and approach M_0 asymptotically according to the exponential expression:

$$M_z(t) = M_0(1 - e^{-t/T_1}). \tag{8}$$

Thus, M_z will achieve 63.2% of its final value (M_0) in a time equal to T_1, 86.5% of M_0 in $2T_1$, and so on. If it is desired to achieve 99% of the total possible magnetization it is necessary to wait for a time of $4.6T_1$. If T_1 is too long, a prohibitively long period can be required to achieve a useful magnetization. The fact that Gorter failed to detect nuclear magnetism in 1936 may have been the result of an unfortunate choice of material, which had too long a value for T_1. It should be noted that there are materials that have T_1 values as long as hours or even days. Fortunately, mobile protons in biological tissues have T_1 values of, at most, a few seconds (Table II).

The only time a signal can be detected from the nuclear spins is when a transverse magnetization is present. This can be achieved by using a short burst of radio-frequency (rf) energy, the B_1 field or rf pulse, at, or very near, the Larmor frequency. A 90° pulse will rotate a magnetization that is initially along the z axis into the transverse plane. If the B_1 field is then turned off, the transverse magnetization $M_t = \sqrt{M_x^2 + M_y^2}$ will precess at the Larmor frequency. Its amplitude will decay according to the relation

$$M_t(t) = M_0 e^{-t/T_2} \cos \omega_0 t. \tag{9}$$

The electric signal picked up during this time is called the free induction decay or FID (Fig. 2). If T_2 is too short, the signal will decay away so rapidly that no useful information can be extracted from the FID. Note that as soon as B_1 disturbs the longitudinal magnetization from its equilibrium value M_0, M_z starts to rebuild according to Eq. (8). Therefore, it is possible to put the spin system through a periodic excitation cycle using a series of 90° pulses. Between pulses the longitudinal magnetization

TABLE II Proton Relaxation Times for Various Substances (37°C)[a]

	T_1 (msec)		
	25 MHz	64 MHz	T_2 (msec)
Pure water (deoxygenated)	4500	4500	2900
Cerebrospinal fluid	4300	4300	2200
Blood (oxygenated)	860	940	350
Skeletal muscle	585	870	47
Brain			
White matter	520	557	72
Cortical gray matter	690	920	100
Liver			
Normal	345	490	43
Hepatoma	810	1080	84
Fat	220	260	84
Multiple sclerosis plaque (mean)	—	1315	174

[a] Adapted from data in Bottomley, *et al.* (1984). *Med. Phys.* **11,** 425–448; Bottomley, *et al.* (1987). *Med. Phys.* **14,** 1–37; Hopkins, *et al.* (1986). *Magn. Reson. Med.* **3,** 303–311; Brooks, and DiChiro, (1987). *Med. Phys.* **14,** 903–913; Hardy, *et al.* (1986). *Magn. Reson. Med.* **3,** 935–940; Larsson, *et al.* (1988). *Magn. Reson. Med.* **7,** 43–55.

rebuilds toward its equilibrium value. At the same time transverse magnetization is precessing and producing a FID signal. Of course, M_z will not have completely achieved its steady-state value M_0 by the time of the subsequent 90° pulse. If the pulses are repeated too rapidly, the Bloch equations show that the magnitude of the FID goes to zero, an effect called saturation. From a quntum mechanical point of view, saturation is the result of an excessive amount of B_1 excitation that eliminates the difference

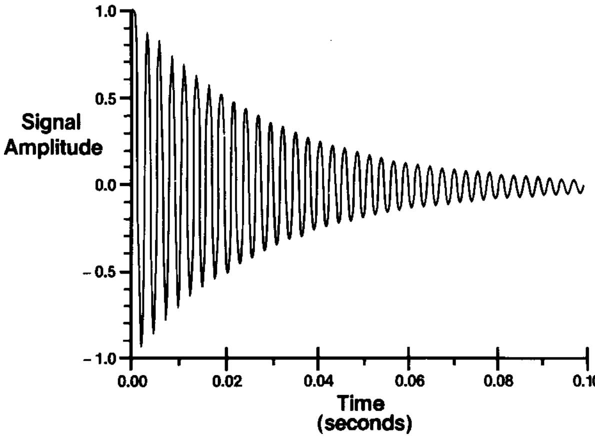


FIGURE 2 Free induction decay. The FID is a damped oscillation. The case illustrated here shows a beat pattern between the signal from spins precessing at the Larmor frequency and a reference frequency shifted from it by 370 Hz. The T_2 is 30 msec. (Courtesy of GE Medical Systems.)

in level populations given by Eq. (1). Once saturation has occurred it is necessary to wait for a time approximately equal to T_1 to permit the magnetization to rebuild [Eq. (8)] before further signals can be obtained from the specimen.

An additional time consideration is the duration of the rf pulse necessary to carry out the desired rotations of the magnetization. The length of this pulse is usually a few milliseconds or less. This is much shorter than T_1 or T_2 for the tissues of interest. Thus, there is no significant change in the length of the magnetization vector during the time that the rf pulse is being applied.

The relaxation times of protons in several human tissues are indicated in Table II. P. A. Bottomley and his coworkers have carried out extensive analysis of the published data, and have concluded that a large variability exists from one investigation to another in published values for relaxation times. Therefore, the values in Table II should not be regarded as precise—standard deviations of ±10% or more in the data are not unusual. It is likely that the improved instrumentation now available, and the recently acquired significance of this data will lead to a rapid improvement in the precision and accuracy with which T_1 and T_2 values for human tissues are known. In general, it appears that for most human tissues T_1 is substantially larger than T_2. For protons T_1 increases in a significant way as the static field strength, and thereby the Larmor frequency is increased. However, T_2 tends to remain constant, or decrease slightly, as the frequency is increased.

As an example of the use of relaxation times to discriminate between normal and malignant tissues, the values for both liver and for a hepatoma, a tumor derived from liver tissue, are given in Table II. In many cases, however, the relaxation time differences between normal and malignant tissues are not so clear cut.

The concept of the T_2 decay is usually explained using the idea of the dephasing of the nuclear spins present in the sample. Consider a transverse magnetization that has been created by a 90° pulse. All the spins in the sample have experienced the same applied magnetic field and immediately after the pulse they all have the same phase. That is, their individual nuclear moments are all pointed in the same direction and the induced voltage they produce in the receiver coil is at its maximum. Once the externally applied B_1 field is removed, however, the individual nuclei are still subjected to the weaker, but persistent, effects of their different local environments. At any instant this leads some nuclei to be precessing faster, and some slower, than the average rate of precession, which is given by the Larmor frequency, set by the external field B_0. The result of this is that the spins get increasingly out of phase with one another as time goes on. This causes the transverse magnetization and, consequently, the induced voltage to decay exponentially toward zero.

It should be realized that the discussion of dephasing provided here is rather casual. Since we are inferring that the dephasing occurs among individual nuclear spins and that this determines the time course of the relaxation of the macroscopic transverse magnetization, the argument should include a quantum mechanical justification. Such an argument would be too lengthy to provide here. This topic is discussed thoroughly in the book by Slichter. For qualitative purposes, the idea of individual nuclear spins gradually dephasing with one another remains a useful one and we will continue to employ it.

Two important physical consequences associated with dephasing are motional narrowing and spin echoes. It might be thought that nuclei that are diffusing rapidly among their neighbors during an FID would experience more rapid dephasing than nuclei that are relatively fixed in position. In fact the opposite is true. The explanation is that if a spin is relatively fixed in position, it will be forced to respond to whatever its local magnetic environment happens to be for an extended period. Unlike the externally applied fields, however, there is no reason for the local magnetic fields to be identical from one location to another because the local environments are not correlated. Therefore, if the sample consists of a large number of localized spins, as in a solid, they will tend to drift rapidly out of phase with one another. On the other hand, if the nuclei are not fixed but move from one location to another rapidly, as in a liquid, the local fields vary from instant to instant, causing the phase of the nuclear precession to increase almost as often as it decreases, and the overall dephasing of the total system proceeds more slowly. Therefore, protons located on fixed sites experience a very rapid dephasing, which means a short T_2, while those on freely diffusing water molecules have much larger values for T_2. After Fourier transformation a larger value for T_2 corresponds to a narrower resonance line. Therefore, this important phenomenon wherein nuclei in highly mobile molecules manifest long T_2 values is called motional narrowing. This remarkable phenomenon was first explained by N. Bloembergen and coworkers in 1948. It is a very strong effect. The T_2 of protons in ice, for example, is shorter than that of protons in water by a factor of 100,000. Therefore, signals from solid tissues, such as teeth and dense bone, decay almost instantly after excitation and cannot be detected by MRI. As a consequence, MRI provides images based on the distribution of mobile protons only. Because the local feilds are random from one nuclear location to another, there is no hope of reversing this form of spin dephasing and it consequent decrease of voltage in the receiver coils.

Another source of dephasing of the spins is caused by macroscopic inhomogeneities in the static applied field B_0. If B_0 varies from one location to another, the spins located in higher fields will precess more rapidly than those in lower fields. The net result is again a dephasing of the spins across the sample and an undesirable decrease of the induced voltage in the receiver coil. However, although the total magnetization of the sample may have decrease to a low value, those spins relatively close to one another in space will still be nearly in phase one another. Thus, although it is not evident externally, a form of spin order still exists within the sample, even after the external signal is no longer detectable. In 1950, E. Hahn showed that this remnant order can be detected by applying a 180° pulse to the precessing spins at a time τ after the 90° pulse that created the transverse magnetization. After such a pulse, the spins that had advanced in phase beyond the average by an amount of $\delta\phi$ are now behind the average by the same amount. Because they are still precessing more rapidly, however, these spins will catch up with the average at a time τ after the 180° pulse. At this time all the spins will be back in phase with one another and the receiver will detect a signal called a spin echo. Of course, all of the microscopic dephasing mechanisms are operating continuously so that the amplitude of the spin echo is reduced by a factor $e^{-2\tau/T_2}$ from the initial amplitude of the FID.

In the early days of NMR, spin echos were used to overcome the relatively high inhomogeneities of the available magnets. In MRI systems today, however, the magnets have sufficient homogeneity that this is not usually necessary. In imaging systems, however, special coils that produce gradients in B_0 are extensively utilized. Spin echoes are often used, as a routine part of the imaging sequence, to reverse the dephasing produced by the deliberately employed gradient fields. If the gradient field is constant in time, a 180° rf pulse may be used to produce an echo as just described. On the other hand, if the gradient field itself is reversed after being applied for at time t, an echo will occur at time $2t$ without the need for a rf pulse. This is called a gradient recalled echo.

The physical environment within biological tissues is extremely heterogenous. Within a single cell there are known to be numerous microscopic structures including the nucleus, the cell membrane, mitochondria, and microtubules. It is likely that the magnetic environment varies somewhat from location to location as a result of this heterogeneity, and consequently, it might be expected that the proton relaxation times would vary with intracellular location. Experimentally, however, it is found that, although the relaxation times vary from one organ to another, there usually does not appear to be a significant deviation from single exponential decay within a single histological region. The explanation involves the rapid self-diffusion of water molecules among one another. At body temperature the self-diffusion coefficient

for water is $D = 3.7 \times 10^{-5}$ cm^2/sec. During a time t the random-walk motion associated with self-diffusion will give each molecule an average displacement equal to $\sqrt{Dt}$. For $t = 10$ msec this gives an average displacement of 6 μm and for $t = 100$ msec, 19 μm. These distances are comparable with average cell sizes. This shows that during a typical T_2 interval a given water molecule has time to encounter much of the intracellular and local extracellular environment, and therefore to average out the inhomogeneities in local field implied by the cellular heterogeneity.

G. Chemical Shift

As discussed, the electrons orbiting each atom respond in a diamagnetic fashion when an external field is applied. As a result, the total magnetic field at the nucleus of an atom is some what less than the externally applied field. The exact value of this reduction depends on the details of the electron orbits in the given molecule and will be different in different molecules. Consequently, in a given external field and for a given type of nucleus, the nuclear resonant frequency will vary from one type of molecular environment to another. This change in frequency δf is called the chemical shift. It is usually small, and $\delta f/f_0$ can be specified in parts per million. These small shifts are easily measured, however, and in MR spectroscopy they are the basis for discriminating one molecular species from another. In MRI the chemical shift leads to an image artifact that slightly confuses the boundaries between fat and water-containing tissues. Protons in water and fat have resonant frequencies differing by 3.5 ppm.

II. NUCLEAR MAGNETIC RESONANCE SCANNERS AS MEDICAL INSTRUMENTS

A. System Considerations

The overall operation of a NMR scanner is controlled by a computer (Fig. 3). It provides the pulse-timing information for the gradient and radio-frequency transmitter coils. It also switches on the preamplifier and the receiver circuitry during the time when data is being acquired from the nuclear spins. The data are acquired as free induction decay or, more commonly, spin echo signals. Extensive mathematical calculations, usually a two-dimensional Fourier transformation, are required to convert the FID or spin echo data into an image and the calculations are done by the computer. For permanent magnet systems no magnet power supply is required; for superconducting magnet systems the magnet power supply is needed only initially when the magnet is being energized.

The NMR signal is sufficiently weak that interfering electrical noise is a major consideration. To minimize outside electrical interference, some form of electrical screening is normally placed around the scanner or the scan room, and electrical filters are used on those circuits that could transmit outside noise to the receiver. Figure 4 shows a patient within a screened room being positioned for scanning. In Fig. 5 the patient is seen within the magnet bore in the location that permits scanning to be done.

The strength of NMR signal generated by the precessing spins is proportional to the degree of alignment of the spins [M_0 in Eq. (2)] and to their rate of precession (γB_0). Since both those quantities are proportional to field

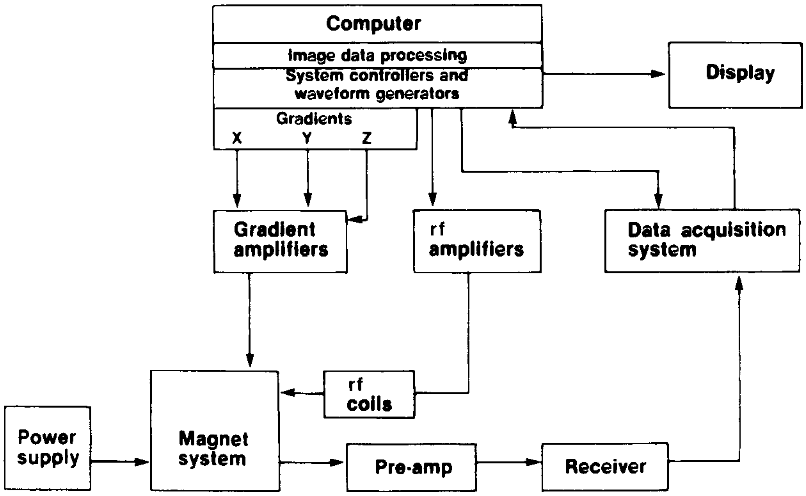


FIGURE 3 Block diagram of an NMR scanner. (Courtesy of GE Medical Systems.)

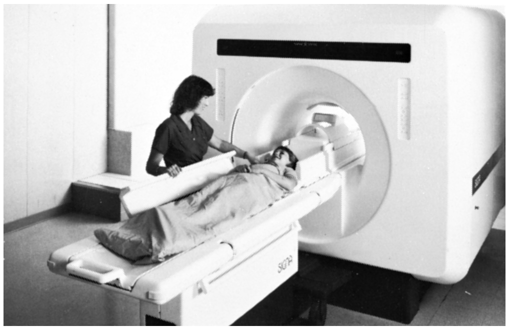


FIGURE 4 Preparation for scanning. The superconducting magnet shown here operates at 1.5 T. (Courtesy of GE Medical Systems.)

strength, the signal increases as B_0^2. The noise in the system can be brought to a very low level by proper design of the electronics. However, there is some electrical noise that is generated within the patient and that cannot be removed by improved circuit design. The ultimate source of this noise is the random motion of ions and charged macromolecules within the patient's body and is of thermal origin. The noise voltage increases approximately linearly with increasing frequency.

A key parameter in determining overall image quality is the signal-to-noise ratio (SNR). The arguments just given indicate that the SNR should increase linearly as the field strength is increased. This is the basis for the use of strong magnetic fields in NMR scanners. If it is desired to achieve high-resolution proton images, thin slices and small picture elements (as will be discussed later) must be used; however, this leads to a decrease in the amount of signal available to determine the pixel brightness numbers. If this process is pushed too far, the images will become grainy because of the effects of the noise. By increasing the field strength of the magnet, the available signal is increased; this can be used to support higher resolution imaging.

FIGURE 5 NMR scan. The patient has been positioned in the center of the magnet and can be seen through a window above the operator's console. (Courtesy of GE Medical Systems.)

In MR spectroscopy, the molecules being studied are present in a concentration very low compared to that of water. As a consequence, spectroscopy signals tend to be very weak. This places an even stronger premium on field strength than does imaging. It is widely accepted that spectroscopic studies of patients are not warranted in field strengths less than about 1.5 T.

B. Magnets

The magnet is probably the most significant portion of a whole-body scanner. The magnets used vary substantially in terms of field strength, shape, and basic design. However, they all must meet certain basic requirements. One obvious requirement is that the magnet be large enough to admit a human body and produce a field strength that is intense enough to produce a strong proton NMR signal. Equally important, though not so obvios, is the requirement for high homogeneity, or uniformity of the magnetic field strength from one part of the imaging volume to another. It is essential that, unless gradient coils are being used to deliberately modulate it, the Larmor precession proceed at the same rate throughout the entire portion of the patient that is to be imaged. The homogeneity requirment puts strong limitations on the acceptable magnet designs. Superconducting and resistive systems are designed to provide a cylindrical symmetry. A cylindrical opening, called the room temperature bore, is available to permit placing the patient into the region of strong and highly homogeneous field.

At the present time clinical proton imaging is usually done using magnetic field strengths in the order of 0.2 to 1.5 T, although, in about 2000, commercial MRI systems operating at 3 T began to become available. Later, in the section on current trends, research scanners operating at even higher field strengths will be described. The types of magnet that have been used to produce clinical images have included permanent magnet systems, iron-core electromagnets, and resistive and superconducting multicoil magnets. Each of these designs has certain advantages. However, for producing fields greater than about 0.5 T, only the superconducting systems are practical; field strengths in this range are beyond the capabilities of present-day permanent magnet materials. In resistive systems the coils are made of a conventional conductor, such as copper or aluminum. If enough current is run through them to produce whole-body sized fields much above 0.2 T, so much heat is generated in the windings that even with water cooling there is a likelihood of burning the insulation or actually melting the wire.

The phenomenon of superconductivity provides the best approach to achieving strong magnetic fields. Since 1911 it has been known that some materials when cooled to

temperatures near absolute zero can conduct electricity with absolutely no electrical resistance and, therefore, no associated heating or energy loss. Electric currents, once started in a loop of superconducting wire, have been shown to persist for years without decreasing measurably, even though no source of electrical energy was used to maintain them. Unfortunately, the original superconducting materials that were discovered, such as lead and tin, could not be used to generate strong magnetic fields because as the current was increased beyond a rather small value the superconducting state was destroyed and the electrical resistance was restored. In the 1960s a new class of superconducting materials capable of carrying much higher current densities was discovered. During the next decade these materials, particularly in the form of niobium–titanium alloys, became the basis of a new class of high-field magnets. These were used to make NMR spectrometers for chemical research applications that were capable of generating much stronger fields than had previously been available. They were only big enough, however, to contain small, test-tube-sized samples. Small-bore, high-field systems are now available for chemical research that provide steady fields in the range of 20–25 T and with NMR proton frequencies approaching 1000 MHz (1 GHz).

By the early 1980s the Oxford Instruments Company of Oxford, England, had produced whole-body superconducting magnets capable of reaching 1.5 T. At the present time several manufactures build whole-body magnets of this type.

In 1986 a still newer class of superconducting materials was discovered, capable of maintaining their superconducting properties to temperatures much higher than the previously known materials. These may eventually have an application in MRI, perhaps by eliminating the need to immerse the coils in liquid helium. The current-handling capabilities of these new materials, however, are at present too weak to permit their use in whole-body magnets. This situation may improve after further research.

The exceptionally high homogeneity of MRI magnets is achieved in two steps: (1) during the basic coil design and (2) by the use of shim coils during operation. The basic approach is to use a set of coils about 1.5 m in diameter positioned along the z axis of the magnet (Fig. 6). The contribution of each coil to the B_0 field is determined by its location along the z axis, its radius, and the number of turns of superconducting wire wound on it multiplied by the current in the coil. The z component of the resulting magnetic field can be represented as an expansion about the center of the magnet by using specific mathematical functions, the spherical harmonics. The zeroth order of this expansion is the perfectly uniform field, $B_z = \text{constant}$, that is the desired field. All other terms in the expansion represent contaminating inhomogeneities

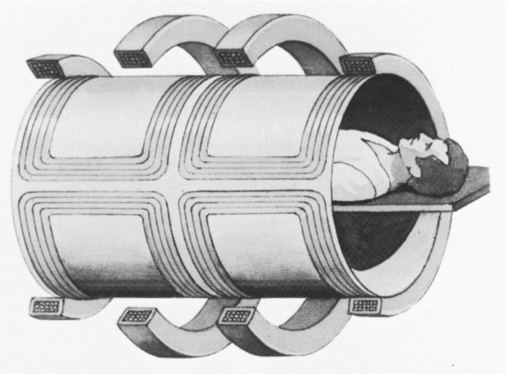


FIGURE 6 Cutaway drawing of magnet geometry. A four-coil geometry is used to produce the static field. The inner coil shown is for the transverse gradient field. (Courtesy of GE Medical Systems.)

that are undesired. The first-order expansion terms represent gradients in all three directions, dB_z/dx, dB_z/dy, and dB_z/dz. These gradients and a large number of additional higher order error terms can be eliminated by correctly placing the proper number of ampere turns at specific locations along the axis. For example, a six-coil design commonly utilized in clinical magnets can theoretically eliminate all the contaminating spherical harmonic terms up to the 12th order. The use of these carefully calculated coil designs greatly increases the volume within the magnet over which the homogeneity specifications can be met.

There are, however, many sources of slight manufacturing errors that prevent the ideal field from being obtained. For example, the individual coils may be slightly out of round, or slightly out of position along the z axis or not oriented absolutely perpendicular to the z axis. To correct for these inevitable manufacturing tolerances, each magnet is equipped with a set of shim coils; up to a dozen or more independent coils are usually available. Each of these coils, wound on a cylindrical coil form near the inner surface of the main field coils, has a different geometry. The geometry of a given shim coil is chosen to produce a field near the magnet center that has a pattern closely approximating a single spherical harmonic. By adjusting the current in each shim coil independently, it is possible to cancel out the residual errors associated with each of the lower order harmonics. The shim coils carry much less current than the main coil windings and, therefore, may be either superconducting or resistive. Sometimes both resistive and superconducting shim coils are provided. The current settings necessary for the shim coils to produce the maximum homogeneity for a given magnet is determined at the time of magnet manufacture in a process called shimming. This process

is repeated episodically, as needed, over the life of the scanner.

Homogeneity is normally specified as the maximum deviation, in parts per million, of the field within a specified diameter spherical volume (dsv) centered on the center of the room-temperature bore. The homogeneity is better, of course, for small volumes. In a typical situation the homogeneity of a shimmed magnet might be 0.1 ppm over a 10-cm dsv, 10 ppm over a 30-cm dsv, and 40 ppm over a 50-cm dsv.

The superconducting property means that such magnets can be operated in the persistent mode. Once the power supply has increased the current to the point where the desired field has been reached (this normally is done over several hours), a superconducting switch is activated, and the power supply is disconnected. As long as the windings are kept below the superconducting transition temperature, no further input of energy is required to maintain the field. Modern magnets, operating in the persistent mode, have no trouble meeting a drift specification of less than 0.1 ppm/hr. This drift rate is so slow that the magnets can go for months or years without requiring additional energy input. To maintain the coils in the superconducting state, they are located within a double cryostat. The inner chamber contains the coils immersed in liquid helium at 4.2 K. The outer chamber contains liquid nitrogen at 77 K as an intermediate temperature reservoir. Every few weeks it is necessary to replace these cryogenic liquids as they boil off.

C. Gradient Coils

In the space between the main magnet coils and the patient it is necessary, for imaging purposes, to place a set of three coils, each of which is designed to produce a specific gradient in B_z, the z component of the static field $\mathbf{B}_0$. These coils are respectively the x, y, and z gradient coils. The currents in these coils are under the control of the computer, and they can be pulsed on and off in the proper sequence to aid in manipulating the spin system as required by the imaging technique. The fields produced by the gradient coils are much smaller than that of the main magnet. The gradient field strengths commonly used in present-day scanners range from 1 to 5 G/cm, which corresponds to 0.0001 to 0.0005 T/cm.

During the imaging session the patient experiences an intermittent series of tapping or banging noises. These sounds can vary in intensity, from rather soft to practically unbearable, depending on the magnitude of the currents involved and on the degree of acoustic damping used. The sounds originate in the gradient coils and result from the magnetic forces between the pulsed gradient currents and the strong, static field.

D. Radio-Frequency Coils

Between the gradient coils and the patient is located at least one additional coil, which serves as transmitter and/or receiver of rf energy. Imaging techniques require a series of 90° and 180° pulses at the Larmor frequency of the protons. These are the B_1 fields used to excite the spin system. The coil that delivers these pulses must be designed to handle the high instantaneous voltages and currents that are required. It should also produce a magnetic field that is as uniform as possible across the desired field of view and at right angles to the static field B_0. To the extent that this B_1 field is not uniform, the pulses produced will produce errors in the desired angles of spin flipping.

The task of building a radio-frequency coil large enough to accommodate the human body, producing a uniform B_1 field, and still capable of resonating at frequencies approaching 100 MHz has provided some technical difficulties. This is because the large size of the coil produces an inherently large inductance, which interacts with the stray self-capacitance of the coil to produce a self-resonance phenomenon that degrades the coil performance. This problem has been overcome by using designs with capacitors distributed along the length of the coil. A particularly effective design for whole-body use at high frequencies has been the "birdcage" concept developed by C. Hayes and his coworkers.

Scanners are usually provided with at least two sizes of cylindrical rf coils. One with a diameter of about 56 cm is large enough to accommodate the entire body. The other, smaller coil is about 28 cm in diameter and is designed for head imaging. Generally speaking, the closer a coil is to the region being imaged, the better SNR it will provide. For high-resolution imaging it has now become common to use coils, called surface coils, that are designed to fit more closely over the region of the body that is to be imaged. Therefore, specialized coils have been developed to image the spine, the neck, the shoulder, and so on.

E. Safety Considerations

In 1976 the U.S. Congress amended the Food, Drug, and Cosmetic Act of 1938 to apply certain restrictions on the introduction of new medical devices. In January 1980 the Food and Drug Administration (FDA) responded to the congressional action by issuing regulations that applied to the manufacturers of new medical devices and to researchers working with such devices. The regulations, analogous to those applied to the introduction of new drugs, made it necessary to develop data regarding the safety and efficacy of new devices prior to seeking approval for marketing them. The NMR scanners were the first major imaging device to be subjected to these

regulations. During the 1980s several manufacturers successfully sought FDA approval for their scanners.

As additional experience was gained during the 1990s, FDA approval was granted for a variety of scanner enhancements, such as the use of higher field strengths and a wider variety of RF and gradient coils.

NMR scanners place the patient in an environment that is quite unlike that of any other medical instrument. Initially, there were several areas of concern that, with experience, have become better understood and appear not to represent a danger to patients. The area of greatest continuing concern is the interaction between the strong, static, magnetic field and ferromagnetic substances inadvertently brought into the region of the scanner. One version of this problem arises from the fringing field surrounding the magnet, which can be treacherously strong. Many objects common in hospitals (e.g., oxygen bottles, mops, fans, and hairpins) contain enough magnetic material that they can be drawn into the magnet with great force and rapidity. Such flying objects are extremely dangerous to anyone in their path. For this reason, most manufacturers and users of the scanners go to great lengths to limit access to the vicinity of the scanner. Permanent magnet systems, and some superconducting systems that have magnetic shielding around them, have smaller fringing fields and are less susceptible to this effect. A second version of this problem comes about because some patients have ferromagnetic substances implanted within their bodies. This is usually the result of a prior surgical procedure, such as the clipping of a diseased blood vessel, but in some cases iron fragments (e.g., shrapnel) have become embedded in a patient's tissues during some sort of accident or explosion. Patients are not always aware of the presence of these objects. Therefore care has to be taken before scanning to exclude those patients with possibly dangerous ferromagnetic implants. Implanted cardiac pacemakers can malfunction, or conceivably, be permanently damaged because of exposure to strong magnetic fields. Therefore patients with these devices in place are not normally candidates for NMR scanning.

Other areas of initial concern were the effects of the static field on normal tissue function; the possibility that electric fields generated by the rapidly changing gradient fields (i.e., the "dB/dt" effect) could cause nerve stimulation or irregularities in the cardiac rhythm; tissue heating associated with the rf excitation field; and possible effects on blood pressure resulting from forces of interaction between blood and the static magnetic field. In the scanners presently used, all of these effects appear to be readily tolerable, and in most cases, negligible. The FDA continues to receive and evaluate designs for more advanced scanners, and the regulatory aspects of scanner safety continue to evolve. Of particular note are guidelines approving the use of static fields up to 4.0 T and, under appropriately controlled conditions, an average heat input to the patient's body (specific absorption rate or SAR) of 4.0 W/kg. The rf heating that occurs under standard imaging conditions is comparable to that resulting from normal metabolic activity and is unnoticeable to the patient. Likewise, the possible effects of dB/dt on nerve excitation and the magnetohydrodynamic interaction between flowing blood and the static magnetic field appear to require conditions far beyond those used in modern scanners. At the present time these concerns represent hypothetical, rather than actual, hazards of the scanning process.

III. IMAGING TECHNIQUES

A. Selective Excitation

One of the key capabilities of MRI scanners is the ability to excite a single, thin slice of spins within the patient. This permits the construction of images that have the character of two-dimensional cross-sectional cuts through the patient's anatomy. NMR imaging resembles CT scanning in producing this type of anatomical image. However, MRI has a substantial advantage over CT in that it permits the location of the slice to be chosen electronically by the operator without moving the patient or any components of the scanner. In MRI, imaging planes of any orientation may be chosen, and these planes may also be moved electronically from side to side, top to bottom, or front to back through the patient's anatomy.

The key to selective excitation is carrying out the rf excitation in the presence of a gradient field. Suppose the static field is uniform across the patient. If a rf pulse is applied at the Larmor frequency, $\omega_0 = \gamma B_0$, for a time long enough to create a 90° pulse, this will excite spins over a large volume of the patient. If, however, the z-gradient coil is used to apply an additional static field, $B_z = G_z z$, at the time of the rf pulse, the resonance condition will be met strictly only in the plane $z = 0$. Spins far from this plane are well off resonance and essentially will be unaffected by the rf pulse. Spins at $z = 0$ will be rotated by 90° just as if no gradient field were present. Spins close to, but not at, the plane $z = 0$ will be partially excited. To predict the exact behavior of the excitation, as a function of z, near the origin it is necessary to carry out a solution to the Bloch equations. It turns out that spins in the selected slice, but slightly off the center plane, will also be flipped through approximately 90° but will have phase differences with those spins at $z = 0$. This will reduce somewhat the signal generated by the slice. It can be shown that a more perfect slice profile will result if, instead of using a rf pulse of constant amplitude, the pulse amplitude is modulated by

an appropriate envelope function. One useful modulation function is $\sin \Omega t / \Omega t$. Here, Ω is an audio frequency that is high enough to permit $\sin \Omega t$ to go through a few cycles during the time the excitation pulse is being applied. Using this or slightly more complicated modulation functions, a rectangular slice profile can be approached. The stronger the gradient applied during excitation, the thinner the resultant slice will be. Typical gradient strengths are of the order of 1 G/cm or less. Typical durations for the excitation pulses are in the range from 1 to 3 msec. The slice thicknesses used in the early NMR scanners were relatively thick—on the order of 10–15 mm. Modern scanners are capable of producing 3- to 5-mm slice thicknesses routinely. The technique just described will excite a slice centered at $z = 0$. Additional audio-frequency modulation of the rf pulse can be used to move the location of the selected slice either up or down along the z axis.

We are using a coordinate system where the z axis points along the patient's body from the head toward the feet, the x axis is horizontal, and the y axis is vertical. The excitation method previously mentioned, which uses a z-gradient coil, will excite slices in the x–y plane; these slices are called axial planes. If, instead of a z gradient, a gradient in the x direction is applied during the rf pulse, the excited plane has a sagittal orientation. The x–z plane excited by a y-gradient field is called the coronal plane. By using a simultaneous combination of x, y, and z gradients various oblique planes may also be excited. Thus by a combination of electronically controlled rf pulses and gradient fields, planes of any orientation and location within the imaging volume may be excited as a first step in the imaging process.

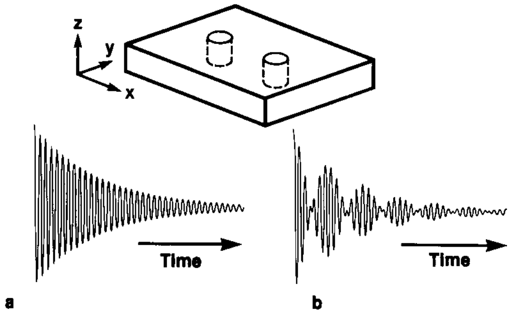


FIGURE 7 Effect of a gradient field on the free induction decay. Water is located in two wells separated along the x axis. In (a) no gradient field is applied and a single damped exponential signal is seen. This is because the protons in both water samples have the same precession frequency. In (b) a gradient is applied along the x direction and a beat pattern is formed between the two frequencies that result. (Courtesy of GE Medical Systems.)

B. Spin-Warp Technique

Several methods for converting NMR data into images have been suggested and demonstrated. The spin-warp technique has found the most wide spread clinical use and will now be described. The two-dimensional image to be formed consists of a large number of individual picture elements, called pixels. There are M rows of pixel elements in one direction and N columns in the other. For mathematical reasons M and N are both usually powers of two. For example, 128×256 and 256×256 are common pixel array sizes. The imaging process must yield a pixel brightness number for each of the $M \times N$ elements in the image. The basic ideas behind the use of gradient fields and Fourier transformation to create position-dependent frequency information are illustrated in Figs. 7 and 8.

The object to be imaged consists of the spins within a slice whose thickness t is determined by a selective excitation process. A desired field of view (FOV) is selected within this slab and is divided into $M \times N$ volume elements called voxels. Normally the FOV is a square. Standard FOV sizes are 24×24 cm for head imaging and 40×40 cm for body imaging. For higher resolution studies of small anatomic regions, smaller FOVs (down to 8×8 cm) can be chosen, usually by varying the strengths of the applied gradients. The brightness to be assigned to each pixel in the image is proportional to the nuclear magnetization in the corresponding voxel. The size of the voxels in the x and y directions are given by $\Delta x = \mathrm{FOV}/N$ and $\Delta y = \mathrm{FOV}/M$. The volume of a voxel is the product of the slice thickness t multiplied by $(\Delta x \Delta y)$.

Figure 9 illustrates a pulse sequence that can produce the data for an axial image. The modulated rf 90° pulse and the simultaneous z gradient are used for the selective excitation of a plane centered at $z = 0$. Immediately after the rf excitation pulse is finished, a gradient pulse in the x direction is used to dephase the spins in the selected

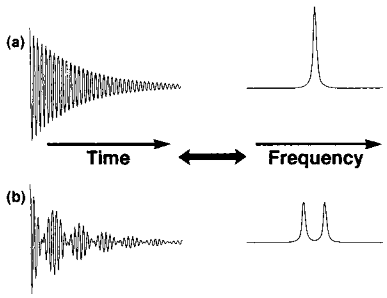


FIGURE 8 Fourier transformation. In (a) the FID is a simple, damped-exponential function of time and its Fourier transform has a single peak at the corresponding frequency. In (b) the FID is a beat pattern consisting of two frequencies and has, correspondingly, two peaks in its transform. The width of a peak is inversely proportional to T_2 that, in this case, is the same for both peaks. (Courtesy of GE Medical Systems.)

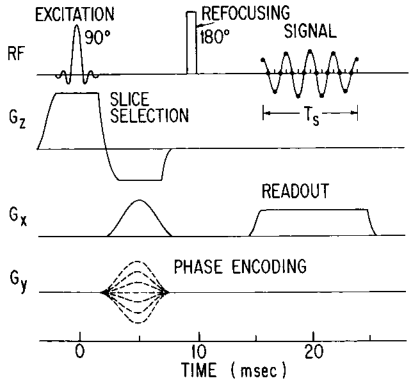


FIGURE 9 Pulse and gradient timing diagram for spinwarp imaging. The top line shows the sequence of events that involve rf signals. The other three lines show the sequence of pulses on the three gradient coils. (Courtesy of Raven Press.)

slice. A 180° pulse is applied at the time t to refocus the spins, and thus a spin echo occurs at the time $2t$. This maneuver permits the separatation in time of the excitation and receive periods and, therefore, the receiver electronics (which deal with a very low-level signals) are not forced to contend with any electronic ringing at the radio-frequency resulting from the very strong transmitter pulse. A pulse of y gradient is also used to give each line of spins at a fixed y position a different phase. The pulse is called the phase, encoding gradient and it is stepped in value each time the pulse sequence is repeated. This generates a different, y-dependent, phase shift for each cycle of the imaging process, and encodes, into the signal, information on the variation in spin density in the y direction. The receiver system is used to detect the voltage in the receiver coil during the sampling time T_s, which is centered on the maximum of the spin echo. A constant x gradient, called the readout gradient G_x^R, is on during the sampling time. This causes the Larmor frequency to vary linearly in the x direction during the time that the signal is being received.

The signal received during T_s is composed of a narrow band of frequencies determined by the readout gradient. A filter is used to limit the detected signals to a bandwidth (BW). The voltage is sampled at N equal intervals during T_s. A criterion due to H. Nyquist states that the bandwidth, the sampling time, and N should be related by

$$\mathrm{BW} \cdot T_s = N. \tag{10}$$

For example, if BW = 32 KHz and $N = 256$, then $T_s =$ 8 msec. The Nyquist criterion assures that if Eq. (10) is satisfied, all the information contained in the signal is also contained in the N digitized sample values. The BW is also related to the FOV by the relation

$$\mathrm{BW} = \gamma G_x^R \mathrm{FOV}/2\pi. \tag{11}$$

By combining Eqs. (10) and (11), the extent of the voxel in the x direction is given by

$$\Delta x = \mathrm{FOV}/N = 2\pi/\left(T_s \gamma G_x^R\right) \tag{12}$$

After a time T_R measured from the beginning of the selective excitation pulse, the process is repeated for a total of M cycles, each of which uses a different value for the phase-encoding gradient. After this process is complete, a $M \times N$ array of digitally sampled data is available in the computer memory. This data can be converted by a two-dimensional Fourier transform technique into $M \times N$ pixel brightness numbers. These numbers can be displayed as an image, which can be viewed either on a cathode ray tube or as a hardcopy on film.

As an example of the voxel sizes used in MRI consider an image of the head using a 24-cm FOV, a 256 × 256 matrix size, and a 5-mm slice thickness. The value of δx and δy will be 240 mm/256 = 0.94 mm. The image will result from the 65,536 voxels in the object each with a volume of 4.7 mm^3.

The phase-encoding process leads to $\Delta y = \mathrm{FOV}/M$. If the sample contains any excited spins that lie outside the FOV in the phase-encoding direction, their signal will be added to the signals from the spins within the FOV and a form of image artifact called aliasing will result. The image is then a type of double exposure, with images of different parts of the anatomy superimposed on one another. If aliasing leads to an unacceptable level of confusion it can be dealt with by increasing M, while keeping the FOV constant (oversampling), and then displaying only the desired portion of the resulting image.

The time between the selective excitation pulse and the center of the sampling interval is called the echo time T_E. Once the FOV and the slice thickness have been selected, the main imaging parameters that can still be varied are T_E and T_R. T_E can be varied between roughly 20 and 200 msec. If T_E is made long, a great deal of T_2 relaxation can occur before the data is taken. In this case, only tissues with long values of T_2 will give strong signals and will appear bright in the image. After each excitation the longitudinal magnetization will start to recover toward M_0. The rate of this recovery is limited by T_1. If the repetition time T_R is short, only those tissues with short values for T_1 can become appreciably magnetized between excitations. Therefore, if it is desired to make a T_1-weighted image, a relatively short T_R is used and T_E is made brief to prevent contrast based on T_2 decay from developing. Conversely, a T_2-weighted image can be created by using a long T_R (up to 2 sec between excitation pulses). This will permit all tissues to magnetize almost fully and eliminate contrast based on T_1 differences. The use of a relatively long T_E will allow differences in T_2 decay rates to become manifest.

The total time required to complete a scan is M times T_R, so that T_2-weighted images generally take longer to acquire. Often it is desired to enhance the SNR by repeating the entire sequence one or more times and averaging the results of corresponding cycles. If there are n repetitions of the basic sequence, the total scan time increases to nMT_R and the SNR is increased by $\sqrt{n}$. The total time to complete an individual scan usually ranges from about 1 to 15 min. Because it is usually necessary to make more than one series of images, the patient is normally in the magnet from 15 to 90 min to complete a diagnostic study.

IV. CLINICAL APPLICATIONS OF MAGNETIC RESONANCE IMAGING

A. Proton Imaging

Once clinical MR scanners became available certain advantages and disadvantages of their use became established. Among the chief advantages of MR are the ability to image almost any region of the body, the very high contrast available between soft tissue structures, the ability to vary the plane of imaging at will, the ability to vary the tissue appearance by varying the scan parameters, the lack of the need for any invasive step such as the injection of contrast agents, and the absence of any ionizing radiation (i.e., X rays). The relative disadvantages include the cost, time requirements, and the inability to detect certain materials. The cost of the examination varies with the time required to complete it, the number of separate images required, and other factors. A large portion of the scanner cost is associated with the magnets, which require expensive materials such as large amount of superconducting wire. Each image acquisition requires a time ranging from less than one second to as much as 20 min. During this time it is necessary for the patient to remain still to avoid blurring the image. CT scans can be taken much more quickly than this. Certain materials (calcium, in particular) that are readily seen in X-ray studies such as CT do not give a NMR signal, and therefore appear only as voids on MR images. This is a drawback especially in the diagnosis of certain tumors. One consequence of this balance of advantages and disadvantages is that CT and MR have become complementary imaging modalities with neither one showing signs of displacing the other.

The advantages of MR scanning have been especially pronounced in imaging of the brain and spinal cord. At the present time a large fraction, perhaps around 70%, of all MR studies are done for central nervous system indications.

Figures 10 and 11, both done on normal volunteers, illustrate several advantages of MRI for brain imaging. Neither the sagittal nor coronal section are routinely available on CT scanners. The contrast in these images is said to be T_1 weighted—that is, the scan repetition time T_R was short enough that not all tissues had time to magnetize completely. Thus, tissues with larger values of T_1, such as the cerebrospinal fluid, appear dark on these scans. Good contrast is seen between gray and white matter of the brain. MR is completely free of any interference from bone. This makes imaging of the cerebellum, near the base of the brain at the back of the head and surrounded by rather dense bone, more effective with MR than with X-ray methods such as CT.

FIGURE 10 Sagittal head image. The image represents a 3-mm thick slice near the midline of a normal volunteer's head. Like all the images shown here this was taken at 1.5 T. (Courtesy of GE Medical Systems.)

FIGURE 11 Coronal head image. The subject was in the same position as in Fig. 10 but by interchanging the gradients used for selective excitation an image of a slice at right angles to that shown in Fig. 10. (Courtesy of GE Medical Systems.)

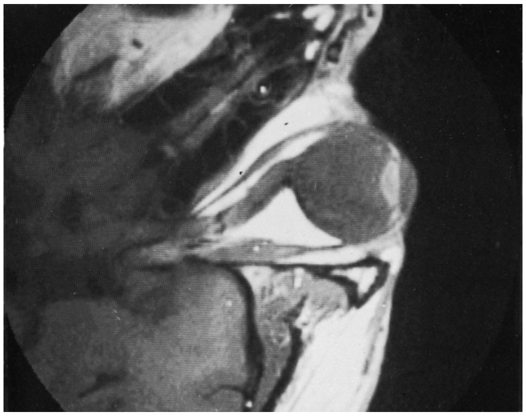


FIGURE 12 Eye image. A surface coil was placed over the right eye to obtain a high-resolution image.

Figure 12 is a high-resolution image of the eye made with a surface coil placed over the right eye. This image demonstrates the capability of MR to image fine anatomical details. The lens and its supporting structures are seen at the front of the globe. The optic nerve exits from the back of the globe and takes a sinuous course toward the brain. The eye is turned to the right and the muscular action to do this is evident. The muscle that turns the right eye to the right, the lateral rectus, is seen short, thick, and contracted. The opposing muscle, the medial rectus, is thin and relaxed.

Examination of joints and the rest of the musculoskeletal system is one of the fastest growing areas of MR scanning. Figure 13 is a surface-coil image of the bones in the region of the right ankle and heel and shows many of the features associated with joint images. The Achilles tendon has few mobile protons, and consequently, shows as a dark band near the left side of the image. It inserts on the heel bone, the calcaneus. Near the center of the image a ligament is seen binding the calcaneus to another bone, the talus. The bright signals from within the bone originate from protons, located in the bone marrow. The cartilage lining the outside of the bone has a grayish appearance in the T_1-weighted image. Experience is accumulating that the details of several of the joints, such as the shoulder, knee, and the temporomandibular joint (TMJ) between the jaw and the skull, are so well seen on MR scans that invasive X-ray studies involving dye injections into the joint spaces can often be replaced.

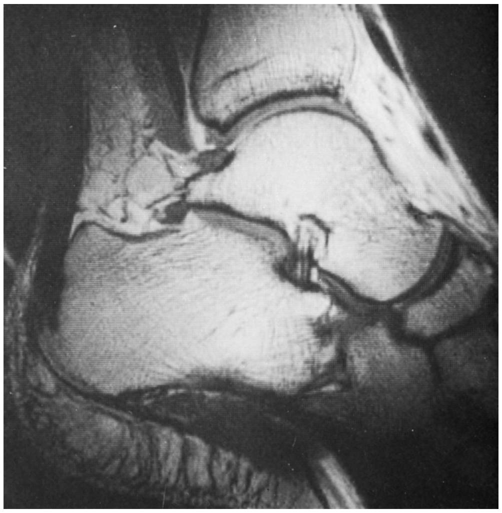


FIGURE 13 Ankle image. A surface coil was used over the right ankle and heel.

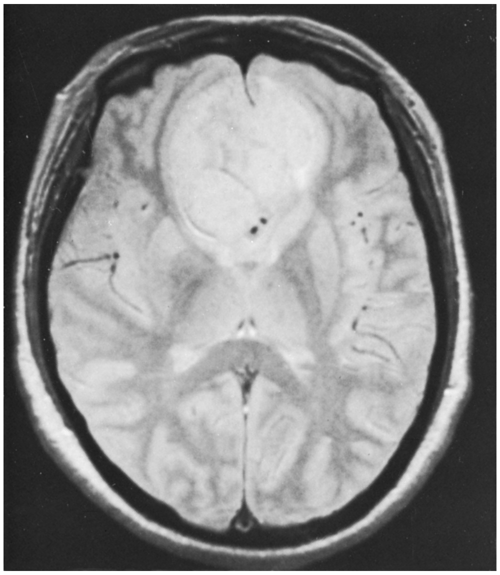


FIGURE 14 Brain tumor. A large tumor, presumably a meningoma, is present in the frontal portion of the brain and is seen on the axial image. (Courtesy of Dr. David Norman.)

Figures 14 and 15 both show large tumors within the brain and illustrate the ability of MR scanners to show sharp demarcation between the normal and abnormal tissues present. Overall, MRI is an excellent method for studying brain tumors because of the good contrast and the availability of multiple scan planes. Neither it nor other modalities, however, can as yet establish a definite diagnosis of the tumor type. MR is also useful in displaying hemorrhage into brain (especially after the initial stages have resolved), strokes, and brain diseases (such as multiple sclerosis), which result from white matter degeneration.

Figure 16 illustrates the power of MR to examine the spinal cord. In this case the patient, with symptoms including weakness and muscle wasting in the hands, has a syringomyelia. This is the presence of a fluid-filled cavity within the spinal cord associated with degeneration of the surrounding tissue. The cavity is seen in this image of the neck as a dark, oblong structure within the cord and

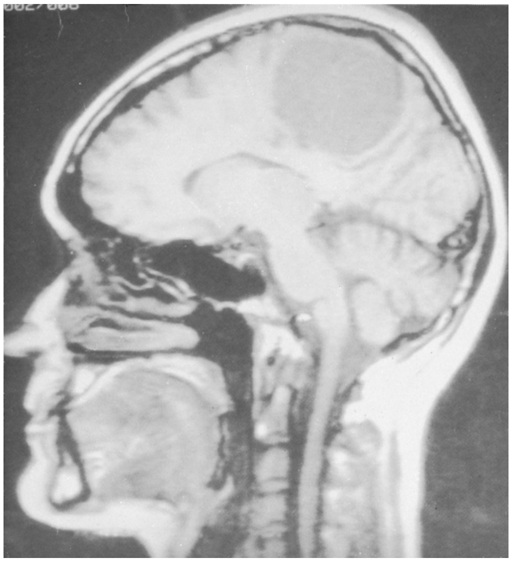


FIGURE 15 Brain tumor. A sagittal image shows a large tumor present in the upper midsection of the patient's brain. (Courtesy of Dr. David Norman.)

FIGURE 16 Syringomyelia. In this disease a fluid-filled cavity forms in the center of the spinal cord. (Courtesy of Dr. David Norman.)

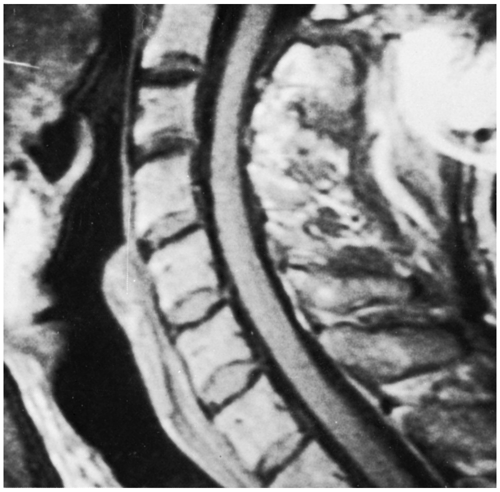


FIGURE 17 Cervical spine image. This high-resolution image of a normal neck is shown for comparison with Fig. 16.

extending almost the length of seven vertebrae. For contrast, Fig. 17 shows a high-resolution image of a normal neck. In this case the spinal cord appears as a solid structure with no evidence of an internal cavity.

Figure 18 is a surface-coil image of the lower spine illustrating disk herniation. Two normal invertebral disks are seen in the upper portion of the image as relatively bright structures. Near the center of the image the abnormal disk, between the fifth lumbar and the first sacral vertebrae, is seen to have herniated, or bulged, backward into the spinal cord. Such herniations are a common cause of nerve injury and irritation often leading to severe low back pain.

B. Current Trends

Magnets in which the patient is placed within a horizontal gap between two magnetic pole faces are now being

FIGURE 18 Herniated disk. This is a surface coil image of the lower spine. (Courtesy of GE Medical Systems.)

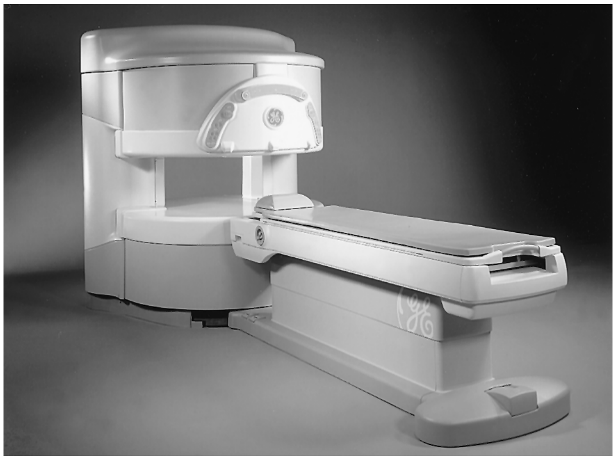


FIGURE 19 High field open magnet. In order to produce field strengths above those achievable with conventional electromagnets, some open scanners, such as this one designed to operate at 0.7 T, utilize superconducting coils to energize the pole faces of the magnet. (Courtesy of Patrick Jarvis, General Electric Medical Systems.)

widely used for the imaging of obese and claustrophobic patients as well as being used as platforms for MR-guided surgical procedures. Physical limitations on the fields that can be obtained with permanent magnets and electromagnets generally limit these scanners to fields less, i.e., less than 0.5 T, than those that can readily be obtained with superconducting cylindrical magnets. Recently the upper field strength limit of these systems has been increased by the use of superconducting coils to energize the magnet pole faces (Fig. 19).

There has recently been substantial activity to develop systems capable of performing image-guided, invasive therapeutic procedures. Because of its excellent ability to provide soft tissue contrast and its potential for very good positional accuracy MRI has a great capability for guiding biopsies and stereotactic surgical procedures. Magnets with either a horizontal or a vertical gap have been designed that allow the members of a surgical team to have direct access to a patient located in the homogeneous magnetic field at the geometric center of the imaging magnet. In such systems the surgeon can operate within a sterile field and interactively control the scan plane and view near real-time images of the operative field on a field-compatible monitor located within the magnet gap. One of the major clinical applications of this technique has been in the area of MR-guided neurosurgery. The advent of MR-guided invasive procedures has created a need for magnetic field compatible surgical instruments and peripheral devices such as electrocardiograms, catheters, and endoscopes.

Since their introduction in the early 1980s scanners using 1.5 tesla superconducting magnets have provided the standard platform for high performance clinical MRI. In the late 1980s a number of research sites began to make use of the improved signal-to-noise ratio available at high field strengths by experimenting with whole-body scanners operating at 4 T. By the end of the 1990s a substantial clinical market began to develop for whole-body clinical scanners operated at fields well above 1.5 T —particularly at fields of 3 and 4 T. This trend was driven initially by the interest of the neuroscience community in blood-oxygen-level-dependent contrast (BOLD) functional MRI (fMRI). This contrast mechanism is associated with the magnetic susceptibility difference between oxygenated and deoxygenated hemoglobin in the cerebral microvasculature, and susceptibility-based contrast is inherently greater at high field strength. The technique of fMRI has provided a revolutionary new capability for the fields of psychology and psychiatry by permitting noninvasive imaging of brain activation by sensory inputs and by thought processes. With the advent of body coil imaging

FIGURE 20 Eight-Tesla whole-body magnet. This scanner was installed at Ohio State University in Columbus Ohio in December 1998 for use in MRI research. It is, at present, the highest field whole-body MRI system. (Courtesy of Dr. Pierre-Marie Robitaille.)

at these field strengths scanners operating at 3–4 T now have the potential of serving as high performance general-purpose clinical scanners.

In addition at a few sites have begun studies using whole-body research scanners operating at fields well above 4 T. For example, in December 1997 an 8 T whole-body scanner was installed at Ohio State University in Columbus, Ohio (Fig. 20). Since that time several 7 T whole-body scanners have been installed and whole body scanners planned to operate in the range of 9–10 T are currently under discussion. Although these magnets have patient bores large enough to admit the entire body, most of the early research applications have been concerned with brain imaging using head coils. Whole-body scanners operating at 7 T and above will probably be used predominately for research, rather than for clinical diagnostic purposes, for several years to come.

SEE ALSO THE FOLLOWING ARTICLES

Electron Spin Resonance • Ferromagnetism • Fourier Series • Magnetic Materials • Microwave Molecular Spectroscopy • Nuclear Magnetic Resonance • Quantum Mechanics

BIBLIOGRAPHY

Abragam, A. (1961). "Principles of Nuclear Magnetism," Oxford University Press (Clarendon), London and New York.

Atlas, S. W., ed. (2001). "Magnetic Resonance Imaging of the Brain and Spine, 3rd ed.," Lippincott, Williams and Wilkins, Philadelphia, PA.

Black, P.-McL., Moriarty, T., Alexander, E., *et al*. (1997). Development and implementation of intraoperative magnetic resonance imaging and its neurosurgical applications. *Neurosurgery* **42,** 831–845.

Bronskill, M. J., and Sprawls, P., eds. (1993). "The Physics of MRI," American Institute of Physics, Woodbury, NY.

Chen, C.-N., and Hoult, D. I. (1989). "Biomedical Magnetic Resonance Technology," Adam Hilger, Bristol and New York.

Edelstein, W. A., Hutchinson, J. M. S., Johnson, G., and Redpath, T. (1980). Spin-warp imaging and applications to human whole-body imaging. *Phys. Med. Biol.* **25,** 751–756.

Fukushima, E., ed. (1989). "NMR in Biomedicine: The Physical Basis," American Institute of Physics, New York.

Hayes, C. E., Edelstein, W. A., Schenck, J. F., Mueller, O. M., and Eash, M. (1985). An efficient highly homogeneous radiofrequency coil for whole-body NMR imaging at 1.5 T. *J. Magn. Reson.* **63,** 622–628.

Hinshaw, W. S., Bottomley, P. A., and Holland, G. N. (1977). Radiographic thin-section image of the human wrist by nuclear magnetic resonance. *Nature* (*London*) **270,** 722–723.

Hollis, D. P. (1987). "Abusing Cancer Science," Strawberry Fields, Chehalis, OR.

Hoult, D. I., and Lauterbur, P. C. (1979). The sensitivity of the zeugmatographic experiment involving human samples. *J. Magn. Reson.* **34,** 425–433.

Jin, J.-M. (1998). "Electromagnetic Analysis and Design in Magnetic Resonance Imaging," CRC Press, Boca Raton, FL.

Kleinfeld, S. (1985). "A Machine Called Indomitable," Times Books, New York.

Lauterbur, P. C. (1973). Image formation by induced local interactions: examples employing nuclear magnetic resonance. *Nature* (*London*) **242,** 190–191.

Ogawa, S., Tank, D. W., Menon, R., *et al.* (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proc. Natl. Acad. Sci. USA* **89,** 5951–5952.

Robitaille, P.-M. L., Warner, R., Jagadeesh, J., *et al.* (1999). Design and assembly of an 8 tesla whole-body MR scanner. *J. Comput. Assist. Tomogr.* **23,** 808–820.

Schenck, J. F. (1996). The role of magnetic susceptibility in magnetic resonance imaging: magnetic field compatibility of the first and second kinds. *Med. Phys.* **23,** 815–850.

Schenck, J. F. (2000). Safety of strong, static magnetic fields. *J. Magn. Reson. Imaging* **12,** 2–19.

Schenck, J. F., Dumoulin, C. L., Redington, R. W., Kressel, H. Y., Elliott, R. T., McDougall, I. L. (1992). Human exposure to 4.0-tesla magnetic fields in a whole-body scanner. *Med. Phys.* **19,** 1089–1098.

Shellock, F. G., ed. (2001). "Magnetic Resonance Procedures; Health Effects and Safety," CRC Press, Boca Raton, FL.

Stark, D. D., and Bradley, W. G., Jr. (1999). "Magnetic Resonance Imaging," 3rd ed., Mosby, St. Louis, MO.

Mass Spectrometry

Kenneth L. Busch
Kennesaw State University

I. Instrumentation
II. Launch Points for the Next Fifty Years

GLOSSARY

Chemical ionization An ion/molecule reaction that leads to ionization of the neutral gas-phase sample molecule via its reaction with an ion generated from the reagent gas present in excess in the ionization source.

Electron ionization The direct ionization of a neutral gas-phase sample molecule via loss of an electron induced by interaction with a high energy (70 eV) electron emitted from a metal filament.

Electron multiplier A transducer for creation of a cascading electron flow initiated by the initial impact of an energetic particle on the front electro-emissive surface.

Electrospray ionization A process through which isolated and usually highly charged ions of sample molecules are created by spraying a fine stream of droplets of sample solution from a highly charged needle, followed by progressive loss of solvent molecules.

Ion trap A mass analyzer that traps all ions from a pulse of ionization at the center of a ring and cap electrode structure energized with DC and radio frequency (rf) fields. Ions of different masses are accelerated out of the trap to an external detector by selective addition of energy.

Matrix-assisted laser desorption ionization A process by which sample molecules cocrystallized with a laser-energy-absorbing matrix are transferred without decomposition directly from the solid into the gas phase by a process of surface ablation, and ionized by proton transfer from the acidic matrix molecules.

Quadrupole mass filter A mass analyzer that uses a structure of four colinear rods of hyperbolic cross section (energized with a combination of DC and rf fields) that creates trajectory oscillations in low-kinetic-energy ions passing through the rods. At a particular set of DC and rf values, only ions of one mass maintain a stable trajectory that allows passage through the mass filter.

Sector mass analyzer A mass analyzer that uses either a magnetic field, or a combination of magnetic and electric fields, to change the direction of travel of high-kinetic energy ions. Ions of different masses are influenced to different degrees, providing the basis for a mass separation.

Time-of-flight mass analyzer A mass analyzer that separates a mixture of ions generated in an ionization pulse by providing all ions with an equal kinetic energy, and then noting the different ion transit times (due to different velocities for ions of different masses) through a flight tube of fixed length.

MASS SPECTROMETRY (MS) today is certifiably distinct from the analytical method of five, ten, and certainly fifty years ago when it was first used as a general purpose analytical tool. Its fundamental principles

remain intact, and the supporting instrumentation continues its steady and innovative evolution. But applications in biological mass spectrometry have expanded at an extraordinary rate, and fundamental new approaches to the creation and interpretation of mass spectrometric information are being developed to address new questions. It has been estimated that a billion mass spectra are recorded daily (Busch, 2000a). The growth and demand for high-throughput mass spectrometry suggests that even that extraordinary number is an underestimate, if not now, then certainly within the next few years. Classically, mass spectra were interpreted to provide details of molecular structure revealed through the unimolecular dissociations of the molecular ion. Increasingly, however, mass spectral data in other forms are being recorded and manipulated. Mass spectrometry is now used to study interactions between complex molecules in the gas phase to provide clues in combinatorial investigations. Mass spectrometric data is transparently summarized in databases that support repetitive high-throughput proteomics analysis. Multidimensional MS/MS data is stored in arrays examined with chemometric-based routines that mine the data in search of answers for complex questions of pattern and structure that have just now begun to be formulated.

This overview covers relevant issues in instrumentation for modern mass spectrometry, an overview of techniques used in both classical and newer areas of mass spectrometric investigation and a preview of growth areas in mass spectrometry research and application. Current applications are best reviewed by examination of the manuscripts published in the current journals of mass spectrometry, and review of the papers presented at the annual professional meetings, especially meetings dealing with biological chemistry, and the annual meetings of the American Society for Mass Spectrometry.

I. INSTRUMENTATION

Francis W. Aston won the Nobel Prize in Chemistry in 1923 for his prolific work in the discovery of isotopes using mass spectrometers that he had constructed in the Cavendish Laboratory. Aston was experienced and skilled in the construction and maintenance and operation of instruments. But, in his 1942 book "Mass Spectra and Isotopes" Aston described his mass spectrometer as an instrument that "behaves at times in the most capricious and unaccountable manner." Where would the field of mass spectrometry be today if this were still the case, if, for example, there had been fewer instrumental developments of reliable and capable ionization sources, mass analyzers, and ion detectors? A well-known adage states that "new instrumentation begets new chemistry"; add a codicil that "new capabilities catalyze new demands," and the intertwined relationship of modern mass spectrometry, its instrumentation, and its market is aptly described.

A. Sample Introduction Systems

The mass analysis and ion detection functions of a mass spectrometer are completed in a vacuum, usually of 10^{-5} to 10^{-6} torr (a slightly higher pressure prevails in the operation of an ion trap mass spectrometer). The processes of electron ionization, chemical ionization, and matrix assisted laser desorption ionization (MALDI) also operate under vacuum, while the electrospray ionization (ESI) source operates at atmospheric pressure, and ESI-generated ions then pass through a set of differential pressure apertures into the mass spectrometer. Mass spectrometers must operate under vacuum so that the ions move through the instrument in a controlled manner rather than being scattered by collisions with residual gas molecules. The mean free path of the ion should be larger than the dimensions of the mass spectrometer itself. The issues of vacuum in mass spectrometry are more thoroughly dealt with elsewhere (Busch, 2000b); the many orders of magnitude difference between instrument operating pressure and the atmospheric pressure of the laboratory, or the pressure regime of a column-based separations method, is a basic design parameter of the sample introduction system.

Mass spectrometry is unique among many modern analytical methods in that sample molecules are physically and irretrievably introduced into the instrument. The sample handling devices must be transport devices that accommodate a wide dynamic range of sample quantities, efficiently transporting all sample molecules from the outside world into the ionization source of the mass spectrometer. Given the diversity of sample types, there is a concurrent diversity of sample introduction systems.

1. Direct Insertion Probe

The direct insertion probe (or direct probe) is a device to introduce small amounts of solid or liquid samples into the ionization source of the mass spectrometer without chromatographic separation. The direct probe is appropriate for use when the sample is known to be pure or of limited complexity, or when a chromatographic separation is impossible because of sample volatility, or when such a separation would be too time-consuming. About 10^{-5} to 10^{-4} torr of sample pressure is sufficient to record a mass spectrum in an electron ionization or chemical ionization source. This sample pressure can be achieved by evaporation of the sample into the vacuum at room (instrument ambient) temperature, or by increasing the temperature of the direct insertion probe, which can be heated

electrically to 400°C. Automation of a direct probe sample introduction system has recently been described (Manura and Manura, 2000). With the automated system, samples can be introduced to the mass spectrometer at the rate of 15–20 samples per hour. This is a faster rate of sample analysis than is normally possible than with chromatography coupled with mass spectrometry, in which the column separation time accounts essentially for all of the time of the analysis.

2. Chromatographic Columns

The ability of mass spectrometry to identify a sample compound is maximized when the sample is pure; the combination of mass spectrometry with a chromatographic inlet system has therefore become a mainstay of instrumental analysis. Chromatography separates sample components in time, passes them into the mass spectrometer, and each is then characterized via measurement and interpretation of its mass spectrum. The first gas chromatography/mass spectrometry (GC/MS) instruments of the middle 1970s used packed columns, a 5-mm outer diameter, 1-m long glass tube filled with the packing material (silica or diatomaceous earth) onto which the stationary phase was coated. An enrichment device was needed to separate the molecules of the sample from the great excess of the helium carrier gas flow. The resolution achieved with such columns was relatively low; peak widths could be tens of seconds wide, and the occurrence of overlapping peaks in the separation of a complex mixture common. If we were still using packed columns for the separation of complex mixtures, mass spectrometry would be limited to analyzing mixtures of 10–20 nonoverlapped components, all exhibiting the proper thermal stability and volatility for GC characterization.

Clearly, the capabilities of modern chromatographic techniques have been vastly improved. Packed column GC has been replaced by capillary column GC. Similarly, the large columns of normal-phase liquid chromatography (LC) are replaced by microcolumn reverse phase LC columns. Capillary electrophoresis (CE) is an entirely new means of separating small amounts of more complex, and charged sample molecules, and has evolved into several distinct forms with unique capabilities. Mass spectrometry coupled with different forms of chromatography is now applied to the analysis of many mixtures, of higher complexity, and more disparate sample types.

It is appropriate here to revisit the meaning of "mixture analysis." A mixture becomes increasingly complex as measured component levels decreases, and as the mixture is examined with increasingly sensitive and sophisticated methods. Specification of mixture components with levels greater than 1% provides a snapshot of composition vastly different from performing the same specification at levels of 0.01%. With increased resolution, chromatography is better able to isolate components present in mixtures at lower levels. With increased sensitivity, mass spectrometry is better able to identify such components. Finally, with increased emphasis on the measurement of analytical information for both screening and regulatory purposes, the demand for trace analyses has increased significantly.

a. Gas chromatography. The combination of gas chromatography with mass spectrometry (GC/MS), itself the topic of an excellent text by McFadden in 1973, is now realized within an integrated, low-cost, widely available analytical instrument. There is no longer any need to review the history of development of GC/MS instrumentation, nor the transition from packed-column to capillary column GC. Instead, it is relevant to consider the simple physical interface between the methods, and then constraints on the operation of the methods in the combined GC/MS instrument, and finally, to preview the information available to the analyst from the GC/MS combination.

In the modern GC/MS instrument, there is no interface *per se* between the capillary column and the ionization source of the mass spectrometer. The influx of helium carrier gas is of such low amount that it can be accommodated without difficulty by the vacuum pump connected to the source, while still maintaining high vacuum in the ionization source and the mass analyzer. The capillary column is terminated by direct connection to the ionization source, with all sample from the column passed directly into the source. Termination of the column at the pressure of the mass spectrometer rather than the usual atmospheric pressure of many other GC detectors does not change the retention time significantly, since the pressure drop occurs in only the last few cm of the column. It is important that a bonded, stable stationary phase be used in the capillary column to minimize the amount of column bleed. Column bleed is elution of the stationary phase itself into the ionization source of the mass spectrometer. Sensitivity is compromised, as is unambiguous identification of compounds due to the presence of extraneous ions in the mass spectrum formed from the stationary phase. Although background subtraction can remove or reduce the contribution of these bleed ions, this requires additional spectral processing time.

As eluting peaks from a GC become narrower (as separation resolution increases), the need for faster scanning of the mass analyzer becomes more stringent. Even for GC peaks only a few seconds wide, at least a few complete mass spectra should be recorded so that they can be averaged together to form an approximation of the mass spectrum measured with a steady sample concentration in the ionization source. Modern mass analyzers, including

sector-based analyzers, quadrupole, and ion trap instruments provide the requisite fast scanning capabilities. No change in the usual operating conditions of the GC is required, and the mass spectrometer is usually operated at near the maximum allowable scan rate consistent with expected ion signal strength.

The usual graphical output of a GC/MS analysis is the TIC trace, where TIC is an acronym for total ion current. The output is graphically similar to the output of a single-channel GC detector such as a flame-ionization detector or an electron-capture detector. In GC/MS, the TIC trace represents the sum of ion intensities across the scanned mass range, a single sum value for each scan number of the GC/MS run. When compounds elute from the GC column, the number of ions formed increases, and the summed TIC value increases. The TIC trace is used to determine the retention times of compounds. Then a complete mass spectrum is assembled by adding together, and then averaging, mass spectra recorded during scans recorded across the eluting peak. Finally, the averaged mass spectrum is searched against the library of electron ionization mass spectra. GC/MS data can also be processed in many other ways, with the reconstructed ion chromatogram a powerful means to extract additional information from the data recorded. The reconstructed ion chromatogram (RIC) is a postrun routine that plots intensities of mass-selected ions vs scan number for a GC/MS run. Ions that "belong" together in the mass spectrum of an eluting compound will maximize in intensity at the same time (the retention time of the compound). Background ions have a steady or slowly changing intensity. Unresolved GC peaks can be discerned by slightly different intensity maxima in the RIC profiles. Data processing in other forms of chromatography/mass spectrometry is similarly accomplished, with the TIC trace providing an overall snapshot of the separation, spectral averaging across the width of the eluting peak generating a characteristic mass spectrum of the compound, and the use of reconstructed ion chromatograms to investigate the purity of the mass spectrum.

b. Liquid chromatography. LC/MS requires an interface and an ionization method that accommodates the polar solvent that carries the separated sample mixture components through the (usually) reverse-phase column. Just as GC evolved from larger packed columns to smaller, higher resolution capillary columns, LC also progressed from columns with large flow rates of solvent to smaller columns that operate with solvent flow rates of a few milliliters per minute, and eventually to microbore columns with flow rates of 20–100 μl/min. Electron and chemical ionization (vide infra) would require that the solvent be completely removed, and that stable gas-phase neutral molecules of the sample be formed. However, LC is used rather than GC specifically because the sample molecules are relatively nonvolatile, and cannot be evaporated without decomposition. Therefore, other ionization methods have been developed that produce ions directly from the solvent (electrospray ionization), or use additives within the solvent to cause a chemical-ionization-like reaction (for example, ammonium acetate buffer was used to cause protonation in the thermospray ionization source).

Today LC/MS usually involves the use of the ESI source, described in detail in the next section. The composition and flow of the solvent is constrained within limits set by stable ESI source operation. A flow rate of 1–10 μl/min of common LC solvents (methanol, ethanol, isopropanol, or acetonitrile) generates a stable ESI spray. ESI interface designs used to accommodate higher LC flow rates include pneumatically assisted electrospray, in which a concurrent flow of warmed gas aids in confining the spray and speeding desolvation, allowing flow rates of 100–200 microliters/minute. Higher flow rate columns can be coupled to an ESI source through a flow splitter. The effluent of lower flow rate capillary columns can be augmented by additional liquid for a stable spray, or one of the newer designs for a nanoelectrospray source can be interfaced to the capillary LC column. Capillary LC columns with an integral spraying tip can also be used for direct electrospray ionization at the end of the LC column, with appropriate connection to a potential source. In all of these interfaces, there is a direct connection between the column and the ionization source, with minimized opportunity for sample loss. Sensitivities are therefore maximized, and are mitigated only by the convolution of the scanning speeds of the mass analyzer and the widths of the LC peaks, as in GC/MS.

c. Capillary electrophoresis. CE was developed in the 1990s, and was soon interfaced to mass spectrometry to meet the need for structure-specific identification of eluting compounds. Often, the sensitivity of the mass spectrometric analysis is mentioned as a driving force for CE/MS coupling, but until development of the latest generation of electrospray ionization sources, laser-induced fluorescence routinely provided lower limits of detection than did mass spectrometry. The potential-driven movement of ions in solution is the basis for CE separations. The rate of movement of ions is determined by the sum of electrophoretic and electroosmotic flow. Both neutral and charged compounds move through the column, migrating at different rates, and maintaining a high separations resolution due to the shape of the flow gradient in the small column. The flow rate through a CE column is a low 1–2 microliters/minute. The sample loading on the CE column is also low, with perhaps a few ng of sample at most available for detection.

The movement of sample through the CE column is driven by a potential difference between the entrance and the exit of the CE column; this applied potential must be taken into account when designing an interface that directs the effluent into the ionization source of the mass spectrometer. CE/MS is exclusively coupled with the ESI source, which has its own requirements for imposition of a high potential on the needle from which the spray emerges. Several early designs were explored for integration of both solvent flow and imposed potentials. Integrated designs now use the exit of the CE column itself as the spraying tip for electrospray. A relevant point in all designs is the fact that the CE column exits at atmospheric pressure, which preserves the pressure profile within the column, and maintains the inherently high CE resolution.

B. Ionization Sources

Mass analyzers cannot manipulate neutral molecules. A positive or negative charge is necessary for interaction of the ion with magnetic and electric fields. The differing magnitudes of responses for ions of different masses is the basis for their separation in the mass analyzer. Most samples exist initially as neutral molecules, and gas chromatography and liquid chromatography are used for the most part to separate mixtures of components that are neutral in the gas phase, and neutral in the liquid used as the LC solvent. The ionization source converts these neutral molecules into ions, or extracts ions from solution, and passes the ions into the mass analyzer of the instrument. The physical means to do so depends on the initial form of the sample. Electron and chemical ionization are the older and more traditional ionization methods, both widely used with GC. Liquid chromatography is used predominantly with electrospray ionization, as is capillary electrophoresis. MALDI is an ionization method that creates ions directly from a solid mixture deposited on a surface (consistent with planar chromatography, but not directly with the column-based chromatographic methods that dominate modern mass spectrometry). MALDI is included here since it is one of the rapidly growing applications areas in biological mass spectrometry, and its use may revitalize some forms of planar chromatography, including forms of planar gel electrophoresis and affinity chromatography.

1. Electron Ionization

Electron ionization was the first ionization method developed for mass spectrometry, and it remains the most widely used. The most extensive mass spectral libraries assembled are those of electron ionization mass spectra recorded under a "standard" set of conditions. The ionization process is the direct result of the interaction of an energetic electron with the electrons in the molecule of interest. The electrons are emitted from a metal (usually rhenium) filament through which 3–4 Amperes of current is passed. This current heats the filament to about 2000°C; electrons are released from the metal and accelerated into the source. The classic "70 eV" electron ionization mass spectrum of an organic compound is obtained when the potential difference maintained between the filament and the source block is 70 V, with the block being maintained at a more positive potential. Variations in the measured mass spectra and in the ionization cross sections of organic compounds with changes in the electron energy were studied early in the development of electron ionization mass spectrometry. A value of 70 V was chosen so that mass spectra recorded at this electron energy did not vary greatly with small changes in the electron energy, and the sensitivity (number of ions produced per amount of sample introduced into the course) is also essentially constant about this value.

The incident electrons must have an energy greater than the ionization energy of the target gas molecule M, defined as the energy required to remove the electron held most weakly within the molecule. The electron ionization process can be written for the gas-phase sample molecule M:

$$\mathrm{M\,(gas)} + \mathrm{e}^-_{\text{filament}} \longrightarrow \mathrm{M}^{+\bullet} + \mathrm{e}^-_{\text{filament}} + \mathrm{e}^-_{\text{molecule}}.$$

Electrons in molecular orbitals are moving at velocities of about 10^8 m/sec. As the very fast filament electron approaches the molecule, it causes the release of one of the slower molecular electrons, forming a positively charged molecular ion. The molecular ion $\mathrm{M}^{+\bullet}$ (the superscripted dot denotes an unpaired odd electron) may subsequently dissociate, since the ionization imparts more energy to the molecule than that required for ionization alone. The excess energy can cause the dissociation of the molecular ion $\mathrm{M}^{+\bullet}$, or it can be retained in the ion as excess internal energy. Since an electron is far too light to transfer kinetic energy to the sample molecule in a collisional process, the process of electron ionization involves electronic excitation of M. The assumption that the atoms do not move as the transition to an excited electronic energy state occurs is known as the Franck–Condon principle. The assumption states, in effect, that the molecular ion $\mathrm{M}^{+\bullet}$ retains the original structure of the molecule M, at least at short times after its formation. If dissociations of the molecular ion are prompt, therefore, we can assume that the dissociations represent those of the original molecule and not a structurally reorganized isomer. Some fraction of the molecular ions formed will be stable enough to pass into the mass analyzer; their measured m/z ratio is a direct indication of the molecular mass of the sample molecule itself. Rationalization of the processes that lead to the fragmentation

processes provides clues to the original structure of the molecule. A balance must be attained between fragmentation extensive enough to provide reasonable clues to structure of the neutral molecule, and a relative intensity of the molecular ion large enough so that the molecular ion can be identified, and the mass of the sample molecule determined. Determination of molecular mass is a physical measurement, in contrast to the insight and intuition of interpretation. The appearance of both $M^{+\bullet}$ and fragment ions derived from the molecular ion—the deduction of both molecular mass and molecular structure—is the core use that has supported the growth of electron ionization mass spectrometry.

2. Chemical Ionization

In electron ionization, an encounter with an energetic electron causes the ejection of a single electron from a gas-phase sample molecule M to form the odd-electron molecular ion $M^{+\bullet}$. If too much energy is deposited into $M^{+\bullet}$ during the ionization process, or if the molecule is prone to dissociate, fragment ions may be seen in the mass spectrum, but the $M^{+\bullet}$ may be reduced to such a low intensity that it is indistinguishable above the background signal level. Without the molecular ion, the determination of molecular mass is difficult. Chemical ionization (CI) was developed to overcome this difficulty and provide molecular ions for compounds used for determination of the molecular mass.

Chemical ionization involves a collision and reaction between an ion and a molecule, both in the gas phase. The ion is called the reagent ion and the molecule is the neutral sample molecule. Since the process is an ion/molecule reaction, parameters such as source temperature and source pressure, sample pressure and reagent gas pressure, and the presence of impurities in the gas stream that enters the source must be carefully controlled. The CI source is a variation of the standard EI source, with modifications required to achieve a higher source pressure (about 1 torr) while keeping the mass analyzer pressure within acceptable limits. The source filament emits electrons, but these are now accelerated to several hundred volts (250–500 eV is typical) compared to the standard 70 eV of an EI source. This higher energy allows the electrons to penetrate through the higher gas pressure in the source. Since the pressure of 1 torr is mostly the reagent gas (often methane), the electron emitted from the filament is likely to encounter a methane molecule. When it does, an electron ionization process occurs, *viz.*

$$CH_4 + e^-_{\text{filament}} \longrightarrow CH_4^{+\bullet} + e^-_{\text{filament}} + e^-_{\text{methane}}.$$

The $CH_4^{+\bullet}$ ion does not travel far before it encounters a neutral gas molecule, and at 1 torr of methane, the molecule it encounters will most likely be a methane molecule. The ion/molecule reaction between the two involves a proton transfer to create CH_5^+ and $CH_3^\bullet$. Several other subsequent reactions occur, and the final distribution of ions depends explicitly on the source temperature and pressure, but the primary reactant ion is usually CH_5^+. This ion acts as a strong gas-phase acid that protonates anything more basic than methane; neutral sample molecules are sufficiently basic to accept a proton to form the protonated molecule. The mass of the $(M + H)^+$ ion reveals the molecular mass, and the protonated molecule then fragments in accordance with the amount of internal energy contained within it. Methane is not the only reagent gas that forms a reagent ion that transfers a proton to the gas-phase sample molecule. Isobutane introduced into the CI source at a pressure of about 1 torr generates $C_4H_9^+$ as an acidic reagent ion, which also transfers a proton to the neutral sample molecule, but gives it less energy so that the protonated molecule fragments less. Ammonia reacts under CI conditions to create NH_4^+. The ammonium ion can transfer a proton to the sample molecule, or it can transfer the intact ammonium group to create $(M + NH_4)^+$. A simple rule in CI mass spectrometry is use methane to obtain the protonated molecule and some fragmentation, and then use isobutane when you want to minimize the fragmentation and concentrate ion current in the protonated molecule (as might be required when low limits of detection are sought). Since the ammonium ion will transfer its proton only to compounds more basic than ammonia itself, selectivity in ionization can be achieved.

3. Electrospray Ionization

Among the ionization methods discussed here, ESI is unique in that it generates ions directly from within a solution that is sprayed from a fine needle at atmospheric pressure. At the core of the ionization source is a stainless steel capillary tube that carries solvent (effluent from a LC column, typically) at a flow rate of 2–5 μl/min. A potential difference of 3000–4000 V is maintained between the needle and a counter electrode, which can be a wall of the source, or a skimmer cone with an aperture that passes the ions into the mass spectrometer. A spray is generated at the tip by the solvent flow emerging at atmospheric pressure, and the potential difference ensures that the droplets emerging from the needle are charged, aiding in their dispersal. A supplemental flow of gas is also sometimes used to aid in desolvation of the droplets. The ions that are within the bulk solvent, or more accurately, those ions that are formed in the droplet as the last evaporation of neutral solvent molecules occurs, are focused by ion lenses, and sampled through apertures, to be passed into the mass spectrometer.

As the solvent emerges from the charged capillary, it first forms a cone (called a Taylor cone) that results as the droplet adopts a shape to mimimize Coulombic repulsions between the charges on the surface of the liquid. The initially formed cone then dissociates into small droplets, each now isolated in the gas phase (and still at atmospheric pressure), and each carrying an excess charge on its surface. Desolvation involves the loss of neutral solvent molecules from the droplet, and proceeds rapidly at atmospheric pressure. As the droplet decreases in size, the charge density increases until an instability limit is reached, and the droplet dissociates into still smaller highly charged droplets. Residual solvent quickly evaporates, leaving only the charged ions themselves to be transferred into the mass spectrometer. The ionic population within a microdroplet, and then a nanodroplet, is not known explicitly. The charge imposed on the droplet as it departs the needle must be carried by a chemical species in the solution. If the droplet is positively charged, the charge carriers will be predominantly protons. As the droplet decreases in size, the "pH" rises exponentially (pH is not an accurate description of the situation since there is no equilibrium). As the droplet decreases in size, the protons are forced to ionize the sample molecules. Protonation, and in fact multiple protonation, is commonly observed.

Key aspects of the ESI process are the formation of multiple charged ions of the sample molecule, and minimum fragmentation of these molecular ions. Positive ions of the general form $(M + nH)^{n+}$ are formed by multiple protonation of larger biomolecules (molecular mass is designated by M) such as peptides and proteins. One effect of multiple charging is to bring multiply charged higher mass molecules within the mass range of commonly used mass spectrometers, since the mass analysis is actually a m/z measurement. Further, since M is constant between the series of peaks observed as adjacent multiply charged ions, the multiple measurements of mass of these ions constitute a series of simultaneous equations that can be solved to determine M, the molecular mass, to a precision of $\pm 0.005\%$. This mass measurement capability was not an improvement on exiting methodology, but was a totally new capability, without precedent and without competition from other analytical methods. This uniqueness makes accurate prediction of future applications and developments in mass spectrometry challenging and exciting.

4. Matrix-Assisted Laser Desorption Ionization

A pulsed laser can be used to deposit a very large amount of energy into a small spot in a short time. This energy causes ionization, and if the desired ions are atomic ions, then laser desorption proves a useful method for creating mass spectra, and even spatially resolved mass spectral maps of a surface. Organic and biological molecules decompose under such intense irradiation, and despite much early work in the area, suitable conditions for successful direct analysis were not found. In MALDI, a matrix is used to moderate on a molecular level the energy deposited at the surface, and promote ionization of the sample molecule without decomposition or excessive dissociation. In practice, the sample is mixed in solution with a large excess of the matrix (typically a smaller molecule chosen for high UV absorbance, appropriate volatility, and ability to promote ionization), and an aliquot of a few microliters volume is deposited on an inert surface to co-crystallize. The surface is irradiated with short (10–20 nsec) pulses of laser light at 337 nm; a tight focus on the surface is not necessary. The irradiance power achieved is approximately 10^6 W/cm^2.

The ions observed in the MALDI mass spectrum (they have been termed the "survivor" ions) are the minority within all the species that leave the surface as a result of the laser beam irradiation, comprising perhaps 0.01% of the desorbed/ablated/expelled species (Zenobi and Knochenmuss, 1999). Since the presence of the matrix is requisite for MALDI, investigators quite logically seek to establish the role of the matrix by changes to its chemical nature, with consequent changes to its intrinsic proton affinity, its absorbance spectrum, its ionization potential, and its crystal structure. While the analyte-to-matrix ratio may be controlled on a macroscopic level, it varies across a broad range on the microscopic level, and will also vary spot to spot. The incident laser beam interrogates a surface on which crystals of different size and different properties are dispersed. The microscopic morphological effects that result from a change in gross sample preparation procedure are uncharacterized. The simple fact of the matter is that MALDI works amazingly well under a wide variety of conditions, and it is perhaps not necessary to control all of them rigidly. It has become recognized that in addition to the direct desorption of species from the laser-irradiated surface, secondary ionization processes occur in the selvedge above the surface. This is a reiteration of the concept of the selvedge introduced to explain reactions in secondary ion mass spectrometry and desorption ionization techniques in general (Cooks and Busch, 1983). Reactions that transfer, preserve, or dissipate charge determine what ions finally survive to form the MALDI mass spectrum. Proton transfer reactions that may involve the matrix form protonated molecules. Alkali cations at the surface as impurities can form cationized species. In a process reminiscent of ESI, multiply charged species can be formed, but the high population of free electrons, and the high mobility of electrons within the selvedge, promotes reduction processes that reform the singly charged ions.

MALDI is carried out in a mass spectrometer that uses a time-of-flight (TOF) mass analyzer (described in the next section). The relevant characteristics of the TOF are its unlimited mass range, its inherent compatibility with a pulsed ionization source, and the fact that all ions of all masses are recorded in the mass spectrum generated by each laser pulse (i.e., there is no scanning of the mass analyzer). The TOF is a high-transmission mass analyzer that allows useful mass spectra to be measured even with the small absolute number of ions that leave the irradiated surface. The relatively low mass resolution of the TOF is rendered irrelevant by the broad isotopic envelope of higher mass ions, and the unlimited mass range (determined simply by how long the analyst waits for the ions to make their way down the flight tube) has moved MALDI-TOF mass spectrometry into the regime of sample molecular masses of 10^4–10^6 Dalton. Finally, TOF is a simple device, and costs are low. Benchtop MALDI-TOF instruments at low cost proliferate into a large number of laboratories, where new uses for the data they provide are found.

The "matrix assistance" in MALDI is a matrix effect like any other. A matrix effect is a deviation in measured response in either direction from that expected from the analyte in the absence of the matrix. The usual assumption is that the deviation is in a negative direction (analytical signal is diminished) or that signal to noise, at least, is reduced in the presence of a matrix. Following this assumption, chemists routinely strive for high sample purity before analysis. However, purification of sample for trace level analysis may leave the sample more reactive or more sensitive to loss processes. MALDI is a superb example in which the matrix supports the ability to perform the analysis in the first place, and the deviation is advantageous.

C. Mass Analyzers

The mass analyzer is the heart of the mass spectrometer. In the mass analyzer, some aspect of ion response to electric or magnetic fields is exploited so that ions of different masses can be differentiated. Salient characteristics of mass analyzers are mass range, mass resolution, ion transmission, and as discussed in a previous section, scan times.

1. Magnetic and Electric Sectors

Ions of different mass but a constant kinetic energy will be dispersed in velocity direction by passage through a magnetic field. The term "sector" derives from a nomenclature that divides a magnetic field into sectors of various angular dimensions. Magnetic sector mass spectrometers are of two general types. Single focusing instruments consist of a single magnetic sector between the ion source and the ion detector, and a double focusing instrument combines an electric sector and a magnetic sector into an analyzer operated as a single unit (Fig. 1).

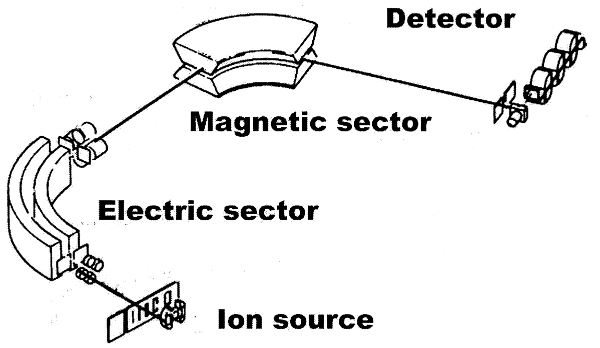


FIGURE 1 General diagram for a double focusing sector mass analyzer.

An ion beam of constant kinetic energy (but containing ions of different masses) is created in the ion source. The kinetic energy is derived from the potential difference (V) between the source and the flight tube of the mass spectrometer, which is usually held at ground potential. Therefore, $zeV = 1/2\ mv^2$, where z is the number of charges on the ion, e is the magnitude of the charge, m is the mass of the ion, and v is the velocity. All parameters in the equation must be expressed in proper mks units (volts, coulombs, kilograms, and meters per second). A charged particle passing through a magnetic field experiences a force that is directed at right angles to both the velocity vector and the magnetic field vector. Centripetal and centrifugal forces are balanced as the ion follows a circular path of radius r while in the magnetic field. Therefore $mv^2/r = zev\mathrm{B}$, where B is the magnetic field strength. Solve both equations for v, and then setting the equations equal to each other yields the classic equation of $m/ze = r^2B^2/2V$. Since e is a constant, it is usually not listed specifically in the equation. If one solves for r instead (the radius of the magnetic field sector), the mathematic expression that results, $r = mv/zeB$, shows explicitly that the radius of curvature of the path of an ion passing through the magnetic field is proportional to its momentum (mv). Since r is fixed, at any given B, only ions with one momentum will pass through the ion optics to make it through to the ion detector of the instrument. Other ions collide with the walls of the flight tube and are neutralized there.

Since r is fixed, there are two parameters that can be varied to change the mass of the ions that passes through the magnetic sector mass analyzer to the ion detector. These are B, the magnetic field strength, and V, the accelerating potential of the source. To maximize the transmission of ions out of the ion source into the mass analyzer, and then out of the mass analyzer into the ion detector, the

components of the mass spectrometer that serve as its ion optics are linked to a set value for V. Changing V to carry out a mass scan will work, but results in a variation in ion transmission known as the defocusing effect. Instead, B is changed in the electromagnet as the current passing through the magnet is changed. Typically, the magnetic mass analyzer is scanned from a higher mass (higher current) to a lower mass (lower current) to allow for maximal heat dissipation. Hysteresis is reduced by allowing a settle time at the beginning of each scan. The reluctance of the electromagnet is a limiting factor in how B can be changed. Additionally, the ions require a finite amount of time to pass through the sector, and the magnetic field must not change during ion transit such that the ion no longer passes through to the ion detector. Scanning speed therefore is an interconnected balance of magnetic field strength, magnet size, accelerating potential, and associated ion optics.

As shown in the derivation above, a magnetic sector alone is a dispersive device based on the momentum of the ions. Proper mass analysis with the magnetic sector requires that all ions leaving the source be provided with precisely the same kinetic energy. The accelerating potential of the source can be regulated to a high precision. However, the ions are formed from neutral molecules in the gas phase, and these have a range of kinetic energies that is carried through their ionization, and then manifest as a range of kinetic energies after acceleration through the source potential. An electric sector (an older term is electrostatic sector) is conjoined with the magnetic sector in a double focusing mass spectrometer. The electric sector consists of two parallel curved plates of specific radius. In the energy focusing achieved by the electric sector, an ion beam (for simplicity, consider all ions to be of the same mass) with a spread of kinetic energies is dispersed by transit through a radial electric field. This dispersion is compensated for by the direction focusing character of the magnetic sector. The electric and magnetic sectors work together as a unit, and therefore double focusing of the ions is achieved. Under such conditions, the exact mass of ions can be measured to a high degree of accuracy using a mass marker (an ion signal of known mass) and a procedure known as peak matching. The voltages applied to the plates of the electric sector do not scan, but rather are set to the values appropriate for the range of ion energies that are to be passed. The mass spectrum is still scanned by varying B, the magnetic field strength.

The mass range of a sector-based mass spectrometer is determined by the strength of the electromagnet, and the value of V, the accelerating potential. Normal mass ranges of several thousand Daltons are typical at full accelerating potential. At reduced V, and therefore reduced ion transmission, higher mass ions can be passed through the mass analyzer. However, for higher mass ions, ESI with a quadrupole or an ion trap mass analyzer, or a MALDI TOF instrument are usually the first instrumental choice.

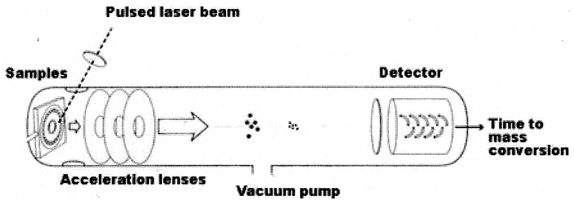


FIGURE 2 Schematic for a TOF mass analyzer.

2. TOF Mass Analyzers

The time-of-flight mass analyzer is a racetrack for ions (Fig. 2). Ions are separated based on velocity differences, and therefore differences in the time required to traverse the length (L) of the flight tube from the ionization source to the ion detector. Ions are accelerated from the source through a potential difference that gives them all the same kinetic energy. But since the ions have different masses, they will have different velocities. Light ions travel faster; heavier ions maintain a more leisurely pace. Solution of the equations show that time through the flight time (t) is related to ion mass: $t = L(m/2zV)^{1/2}$. A short pulse of ions enter the flight tube at the same time, and then disperse along the path of ion movement over time. Measuring the time of arrival of the ions determines their mass.

The TOF mass analyzer is based on an instrumental development that first came to fruition in the 1950s, but then experienced a renaissance in the 1990s, bolstered both by new technology (the ability to measure ion arrival in progressively narrower windows), the ability to compensate for initial ion energy spread using an energy focusing ion mirror (Mamyrin, 1994), and the need for a mass analyzer to complement the abilities of pulsed ionization exemplified by MALDI. Synergistically, the demand for higher mass capabilities, and a simplified approach to mass measurement, in biological analysis developed at the same time. The development of the TOF mass analyzer has literally pushed mass spectrometry into totally new areas of application. Ion masses are routinely described in units of kDa (kiloDaltons), and with higher resolution attained in sophisticated instruments, these centroid masses are measured with accuracy and precision (about 0.1% mass measurement accuracy). Even more amazing is the demonstrated sensitivity; a few femtomoles of sample suffices for the determination of mass. Issues for the future relate to the ability to control and physically manipulate such small amounts of sample outside of the mass spectrometer. The analyst may engineer devices and processes to accomplish this goal. Nature precedes us; a single cell is a smartly packaged collection of sample molecules.

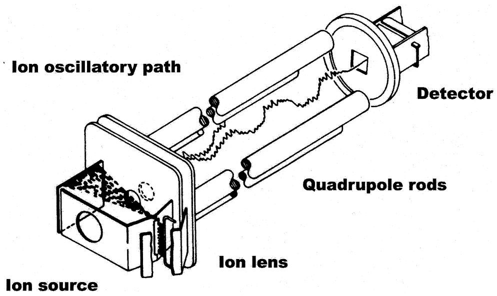


FIGURE 3 Schematic for a quadrupole mass filter.

Single-cell analysis by MALDI-TOF, and the identification of bacteria through direct MALDI analysis, have been reported by several research groups.

3. Quadrupole Mass Filters

The quadrupole mass filter (Fig. 3) is a widely used mass analyzer that has benefited from 30 years of continual commercial development and refinement. It consists of four accurately aligned parallel metal rods that are arranged symmetrically around a central axis that is the path of ion movement from ion source to ion detector. An electrical field is created around the central axis by DC and radio frequency (rf) potentials placed on the rods, with opposite rods connected together. The electrical signals placed on the rods determine the paths that low-kinetic-energy ions follow through the rod structure. At a given DC potential and rf frequency, only ions within a certain mass range follow a stable ion trajectory that allows them to reach the detector rather than collide with the rods themselves. The width of that range is electronically adjustable, and is usually set to be 1 Da in width. The oscillating field applied to the rods alternately attracts and repels ions passing through the mass filter, inducing an ion motion that is exploited to differentiate ions on the basis of their mass. The mathematical equation that describes the motion of the ions in this field is known as the Mathieu equation. The reduced form of the Mathieu equation can be presented graphically in what is called a stability diagram, which shows both the several stability regions for ion trajectories within a quadrupole mass filter, based on the values of the electronic parameters, and the scan function that is followed to pass ions of successively different mass through the structure to the ion detector. The scan is accomplished by scanning across a range of values for both the DC potential and the amplitude of the rf potential, while keeping frequency and the ratio between them constant (ref).

The initial advantage (1970s) of the quadrupole mass filter for GC/MS was its faster scan speed than sector-based instrumentation. In addition, the small physical size of the quadrupole mass filter (15–20 cm in length) resulted in a smaller and less expensive instrument package. Continued refinement culminated in the benchtop GC/MS instrument, a major step forward in making mass spectrometric capabilities more widely available. The limited mass range of the quadrupole mass filter (initially 1000 Da) was not an impediment in GC/MS work, and adequate sensitivity could be obtained with the usual electron and chemical ionization sources. Ultimately, quadrupole mass filters were marketed with upper mass limits of 4000 Da; this extension of the upper mass limit required physical changes in length and the radius of the rods, and changes to the driving electronics as well. Uses in higher mass analysis are now accomplished via formation of multiply charged ions in an ESI source. In these applications, the quadrupole mass analyzer is in competition with the ion trap (described in the next section).

4. Ion Traps

Commercial instruments based on the quadrupole ion trap (QIT) were first shipped in 1984, although the basic device and the basic principles for its operation were first described in 1953. In contrast to beam instruments in which ions physically move in a path from ion source to the mass analyzer to the detector, the QIT is an instrument in which the ions remain essentially in one place (the center of the trap), and are manipulated in a time sequence. The QIT (Fig. 4) is constructed of three electrodes. (The term "quadrupole," then, can be confusing, but emphasizes the relationship of this device to the quadrupole mass filter.) The doughnut-shaped central (ring) electrode is sandwiched between two end-cap electrodes. In the simplest incarnation of the QIT, a gated electron beam from a filament enters through a small aperture in one of the end-cap electrodes to cause ionization of gas-phase neutral sample molecules resident in the central portion of the device. The other end-cap electrode also has an aperture through which selected ions reach the electron-multiplier detector. All electrodes present a hyberbolically curved surface to the central cavity.

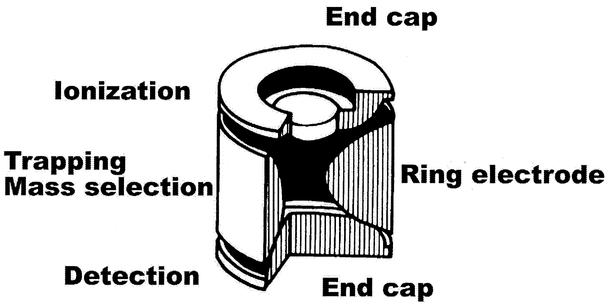


FIGURE 4 General diagram for an ion trap mass analyzer.

Once the ions are formed through a process of electron ionization, they are trapped within the QIT by application of appropriate voltages (DC) and rf signals on the electrodes. Again, in a simple case, the DC voltages can be at ground potential. A rf drive potential of about 1 MHz frequency is applied to the ring. Ions within a broad mass range trace stable orbits near the center of the QIT. In most instruments, helium gas at a pressure of approximately 10^{-3} torr aids in keeping the ions in orbits near the center of the trap. As the amplitude of the rf drive potential is increased, the ion motions progressively change. Eventually, ions develop an unstable trajectory along the z-axis of symmetry (see figure), and pass through the ion exit holes in the end-cap electrode to be detected by the electron multiplier. Appropriate potentials are used so that ions are ejected from their orbits in mass order, and thus the device scans across the mass range of the mass spectrum.

5. Fourier Transform Mass Spectrometers

The basis of high resolution Fourier Transform mass spectrometry (FTMS) is the measurement of the frequency of an ion orbiting in a static magnetic field. The ions are not destroyed by this measurement of frequency. Instead, the ions orbit continuously within the confining cell of the FTMS instrument (held there by a combination of potentials applied to a cubic cell and a static magnetic field) until removed by application of an external electrical pulse, or scattered out of a stable orbit by collisions with neutral residual gas molecules. The key to an accurate measurement of frequency, and therefore an accurate determination of mass, is maintenance of a coherent ion orbit. With extended measurement time, the impact of random errors in the determined frequency value are reduced (the usual measurement statistic in which random noise decreases with the square of the number of measurements). The keys to the maintenance of ion orbit are the homogeneity of the static magnetic field, the symmetry of the trapping potentials generated within the ion confinement cell, and the base pressure within the FTMS instrument. At low base pressures within the FTMS instrument, the ion orbits are stable for periods of seconds and longer. Extraordinary resolutions have been achieved not only for simple atomic ions but also recently for complex ions derived from complex organic and biomolecules, as well as for ions from polymers and organometallic compounds. The maximum resolution of 100,000 reached by sector instruments has been surpassed easily by a resolution of 1,000,000 readily achievable in FTMS.

Perhaps the most significant practical impediment to the use of FTMS has been the need to maintain a low pressure in the cell where the ions orbit. A pressure of 10^{-9} torr is desirable, as compared to the 10^{-6} torr that can be accommodated in other mass analyzers. This very low pressure can be reached, but engenders additional difficulty in creating and transporting ions from the outside world into the cell. The interfacing of an ESI source to a FTMS has been accomplished, but the practical difficulty of transporting ions efficiently through a pressure differential of 10–12 orders of magnitude is considerable. The advantages of high-resolution mass spectral data are sufficient impetus for development of this instrument and its expanding applications. The presence of multiple charges on an ESI-generated ion increases the mass range of FTMS by, in effect, making the ion easier to manipulate in the cell and easier to detect. In that the ion is not destroyed in its detection, the sensitivity of FTMS is high. The limiting steps are maintaining charge on the ions (larger ions have a tendency to relinquish their charge unimolecularly) and maintaining those ions within a stable and coherent orbit. With enough time, the signal frequency can be determined for a group of only 100–1000 ions of the same mass, which may correspond to only a very small amount of sample, or a very low abundance ion in a mass spectrum. In fact, the presence of too many ions complicates the measurement through introduction of secondary space charge effects that distort the measured frequency. Capabilities of FTMS are discussed again in the final section of this overview on launch points for the next fifty years.

D. Detectors

With the exception of the Fourier transform mass spectrometer described above, ion detection in other mass spectrometers is the destructive event in the sequence. Each individual ion carries such a small charge and is of such low absolute mass that direct detection is difficult (but not impossible). Considerable amplification of the ion charge or the ion mass is necessary to make mass spectrometry practicable. Electron multiplier detectors, known since the 1950s, provide the requisite amplification of the charge into an easily manipulated current.

1. Electron Multipliers

The electron multiplier detector is a transducer that converts the impact of ions on its front surface into an amplified electron current then sampled by modern digital electronics. The sensitivity of mass spectrometric analysis is supported by the high gain (as high as a millionfold) provided by modern electron multiplier. This impressive gain is achieved by a combination of the electron emissive properties of the active surface of the multiplier and the potential difference maintained along its length. The active surface of the modern Channeltron version of the electron multiplier consists of emissive layers of silicon dioxide overlying a conductive layer of lead oxide supported by

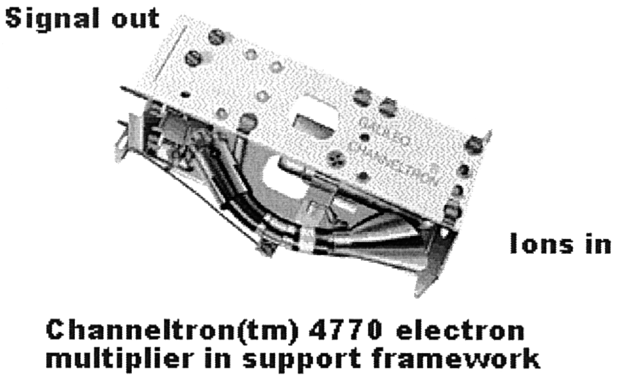


FIGURE 5 Diagram for a continuous dynode electron multiplier.

a bulk glass substrate. The supporting glass structure is stretched into a tube of characteristic shape in a "continuous dynode" electron multiplier (Fig. 5), with a defined resistive path between the front of the device and the output at the end. In contrast, a discrete dynode electron multiplier is composed of 12–20 metal surfaces (dynodes composed of a copper-beryllium alloy oxide) connected in series through discrete, vacuum-compatible resistors. Electrons move in paths that intersect consecutive dynodes along the series, or along the tube, in both types of electron multipliers.

The relevant characteristic of the active surface in any electron multiplier is the secondary emission coefficient. This value, coupled with the potential gradient maintained from the front to the back of the device, and the length-to-diameter ratio of the tube (in Channeltron multipliers) determines the gain that can be achieved, along with the parameters of operation. Assume that a positive ion has been passed through the mass analyzer and approaches the front of the electron multiplier. A −2000 V potential is applied to the front of the electron multiplier, and the output of the electron multiplier is referenced to ground. The positive ion impacts the multiplier active surface with a final kinetic energy determined by the −2000 V. If the ion has passed through a quadrupole mass filter, it is accelerated from the relatively low kinetic energy maintained during mass analysis to a higher velocity. If the ion has passed through a sector mass spectrometer (with a relatively high kinetic energy), it is usually decelerated before impact with the front of the multiplier. The velocity of the impacting ion must exceed the threshold required to cause the emission of electrons from the active surface on ion impact. The first step in the operation of the multiplier is therefore the transformation of a primary positive ion impact into a release of electrons from the active surface. The ion impact releases several electrons from the specially prepared active surface, and the released electrons are accelerated to a more positive potential within the device, whether it is the next discrete dynode, or further down the tube of a continuous dynode device. The electrons acquire a kinetic energy equal to the potential difference between their point of origin and their next point of collision with the surface. Electrons typically gain a few tens of eV of energy in each transition, sufficient energy to cause the release of several more electrons. Two electrons initially released by positive ion impact generate 4 electrons at the second impact and release event. As the process is repeated, the four electrons become 8, and so on. A cascading effect is established such that each incident particle at the front of the device produces an amplified current of electrons at the output. The gain of the device is ultimately limited by the space charge that accumulates within the amplification channel of the device, as this disrupts the progressive travel of electrons to the more positive surfaces. The electron multiplier current is directed through a vacuum feedthrough to a low-noise preamplifier, and then to an amplifier. Between these two stages of amplification, several additional orders of magnitude of gain are achieved. The current is transformed into a voltage (usually in the range of microvolts to millivolts), then sampled by the analog-to-digital converter (ADC) and recorded by the data system. The entire process of amplification and digitization occurs rapidly (within a few microseconds) so that the amplitude of the ion signal is recorded in the appropriate mass window.

For positive ion detection, the front dynode surface of the electron multiplier is maintained at a high negative potential to attract the positive ions, and the output signal is referenced to ground. For negative ion detection, the converse is true. The first active surface must be held at a high positive potential (+2000 V) to attract the incoming negative ion. But the successive surfaces of the multiplier must be still more positive to attract the emitted electrons. The same 2000 V potential gradient across the multiplier means that the current output will now be carried on a +4000 V reference from which the signal must be decoupled. Although there are electronic means of decoupling (often involving a photoconversion step), a more practical solution to the detection of negative ions involves the use of a separate conversion dynode within the electron multiplier assembly. The conversion dynode is separate from the main body of the multiplier, and can be held at an independent potential. For the detection of negative ions, a high positive potential is imposed on the conversion dynode. Negative ions are accelerated toward and collide with the surface of the conversion dynode, which is composed of materials chosen so that the collision causes the release of electrons, positive ions, and photons. Depending on the relative potentials of the conversion dynode and the front of the multiplier, either positive ions or electrons can be collected at the front of the multiplier; usually positive

ions are collected. Once these positive ions are accelerated into the body of the electron multiplier, the electron release and amplification process occurs as described above. For the efficient detection of high-mass ions, the conversion dynode may be held at a very high potential to accelerate larger ions to a velocity sufficient to release on impact a large number of lower-mass positive ions that are then detected by the electron multiplier.

2. Array Detectors

The electron multiplier is a single channel device. The mass analyzer selects the ions by mass, and the ion current is generated through amplification as described above, measured, and recorded as the value for the ions at that mass. Since the mass analyzer has to scan through the mass range, at any given instant, all ions except those of the one mass passing through to the detector are lost. The Mattauch–Herzog geometry of double focusing mass spectrometer used a focal plane detector in which ions of different masses were brought simultaneously into focus at a planar detector that most often was photographic film. The sensitivity of this integrating detector was high. With newer developments in multichannel or multipoint collectors and their associated electronics, photographic film has been replaced with array detectors, which are assemblages of small devices that act individually much like the electron multiplier described above.

The term array is used to describe an assemblage of small single point ion detectors (sometimes called elements) arrayed in a plane. Each of these elements acts as an ion current amplifier. Each element acts as a detector for the ion mass that is directed onto that particular space, and that particular element. All ions of all masses are detected simultaneously, and the array detector acts as an integrating detector (ignoring the electronics of readout and measurement, and transfer of values into the data system). Clearly the array detector will provide an increase in sensitivity over a single-channel detector. It is therefore useful in trace analysis, or in analyses in which a fraction of the mass spectrum (perhaps across the molecular ion region) should be recorded without scanning that cover a wide mass range. More recently, these detectors have become useful in mass spectrometers in which the production of ions in the source is discontinuous, as in a laser desorption ionization source. Time variabilities in ion production are muted by the integrating nature of the detector.

II. LAUNCH POINTS FOR THE NEXT FIFTY YEARS

The mass spectrometer represents our microscope into the world of individual ions, recording their physical attributes such as exact mass and measuring their intensity as formed from a dissociating molecule. In interpreting a mass spectrum, we explore issues of chemical reactivity, often based on kinetics, but expanding more recently into thermochemical arenas. This chemical focus represents the fifty years of modern mass spectrometry in contrast to Aston's world of isotopic discovery that dates from 1919. Prognosticators look to the next fifty years, lest the content of an overview such as this become archaic even as it appears.

A. Exact Mass Measurement

Since the exact masses of individual atomic ions are known with high accuracy, the exact mass of an ion of a given empirical formula (a known combination of atoms) is also known with high accuracy by simple summation. In mass spectrometry, measurement of the exact mass of an ion is used to deduce information about the empirical formula (*not* the molecular formula) of the ion. If the total number of possible combinations is small, and the measurement of the exact mass sufficiently accurate, the measurement can be used to derive an ion empirical formula. The traditional exact mass measurement is usually limited to the more abundant ions observed in the mass spectrum, since higher mass resolution in sector instruments was achieved at the cost of lower ion signal.

High-performance sector instruments could provide mass resolutions of up to 100,000, but the typical daily operation was a resolution of a few thousand. Excursions to higher resolutions required concerted efforts with a clean and stable instrument, and the availability of a relatively large amount of sample. As described above, the basis of high resolution FTMS is the measurement of the frequency of an ion orbiting in a static magnetic field. The ions are not destroyed by this measurement; the measurement can be completed on only a few hundred ions kept within the cell for an extended period.

The ability to routinely make exact mass measurements as exemplified by FTMS is not a simple extension of the use of such values to derive empirical formula. There are two related areas in which the mass measurement data is used in situations in which other information provides additional restraints on possible empirical formulas, and a synergistic extension of abilities. Rodgers *et al.* (2000) show that stable isotope incorporation into selected biomolecules increases the upper mass limit at which accurate mass measurement can provide an empirical formula composition for the ion. Simply, the nominal mass difference between a "natural abundance ion" and the corresponding ^{13}C-enriched (99% enrichment) ion of the same form immediately yields the number of carbon atoms in the molecule. Once that value is known, the

number of possible empirical formulas that can sum to the accurately measured ion mass is greatly reduced. The approach is general to any particular isotopic incorporation. It is not a development in instrumental capability; it is a development in which instrumental capability is synergistically coupled with reasoned use of other information. The same coupling is evident in the ability of FTMS to provide accurate mass measurements for ions generated in multistep dissociations in MS/MS. It is not necessary that the value of exact mass measurement differentiate between all possible ion empirical formulas, but only between possible empirical formulas for product ions that can be formed from the mass-selected parent ion for which the empirical formula is already known. As a simple example, if the parent ion contains no sulfur atoms, then the product ion cannot either. Further, "weak" points within a molecular structure, and therefore probable sites of cleavage, are usually apparent to experienced analysts. With the number of possibilities thus even further reduced, the value of the exact mass measurement is amplified further. The possibilities narrow as the number of stages of induced dissociation increases, relaxing the need for accurate mass measurement as the ion signal grows smaller.

B. Multidimensional MS/MS

The central analytical role of mass spectrometry continues to be twofold: what is the sample and how much of it is there? Certainly there are extraordinary new insights into chemical behavior between complex species in solution and in the gas phase that will be part of the next fifty years of mass spectrometry, made possible by advances in ESI and MALDI. Forecasting in the new realms of biological mass spectrometry is difficult. But in looking exclusively to these new and exciting areas, the implications of modern mass spectrometry for the "tried and true" are often overlooked. In the introduction, it was stated that a billion mass spectra are recorded daily around the world. Most certainly, the vast majority of these are never examined by human hand and mind, neither assessed nor interpreted. Some decision of some value is reached, perhaps automatically, and the mass spectra are then archived. With tremendous advancements in computing and data storage capabilities, we now follow the same process with MS/MS data. At the same time, new instrumental capabilities (specifically with ion traps and Fourier transform mass spectrometers) mean that multiple-stage MS/MS can be completed, and the sequential induced dissociations of a mass-selected parent ion into product ions can be measured. It is easy to record and easy to store multidimensional MS/MS data.

What does this data mean? It is most certainly a series of insights into the structure and reactivity of the ions. But in a larger sense, it is also a collection of a pattern of information that can reach a complexity sufficient that the pattern itself becomes a specific identifier for a particular molecular structure and identity. This concept is not unfamiliar to users of nuclear magnetic resonance (NMR) data. Targeted sifting of a complex database of multidimensional MS/MS has not yet impacted the analytical community, and the clear advantages for unambiguous sample identification have not yet been reaped. It is merely a matter of time, and closer to five years than fifty.

SEE ALSO THE FOLLOWING ARTICLES

Analytical Chemistry • Databases • Gas Chromatography • Ion Kinetics and Energetics • Liquid Chromatography • Mass Spectrometry in Forensic Science

BIBLIOGRAPHY

Aston, K. W. (1942). "Mass Spectra and Isotopes," Arnold, London.

Busch, K. L. (2000a). *Spectroscopy* **15**(11), 30–39.

Busch, K. L. (2000b). *Spectroscopy* **15**(9), 22–25.

Cooks, R. G., and Busch, K. L. (1983). *Int. J. Mass Spectrom. Ion Phys.* **53,** 323.

Mamyrin, B. A. (1994). *Int. J. Mass Spectrom. Ion Phys.* **131,** 1–19.

Manura, J. J., and Manura, D. J. (2000). *American Laboratory* **33**(3), 40–53.

McFadden, W. (1973). "Techniques of Combined Gas Chromatography/Mass Spectroscopy: Applications in Organic Analysis," John Wiley Interscience, New York.

Rodgers, R. P., Blumer, E. N., Hendrickson, C. L., and Marshall, A. G. (2000). *J. Amer. Soc. Mass Spectrom.* **11,** 835–840.

Zenobi, R., and Knochenmuss, R. (1999). *Mass Spectrom. Rev.* **17**(5), 337–366.

Mass Spectrometry in Forensic Science

Jan Schuberth
National Laboratory of Forensic Chemistry, University Hospital (Emeritus)

I. Data Generation
II. Data Evaluation

GLOSSARY

Analyte The target substance in a sample that is searched for and identified by chemical analytical means.

Chemical ionization A low-energy mode of ionization by which a reagent gas, after having been ionized by electron impact, reacts with the analyte also in the gas phase. By causing only minor fragmentation of the analyte, this soft ionization often results in a charged analyte molecule.

Electron ionization A high-energy mode of ionization by which electrons are allowed to bombard an analyte molecule in the gas phase, whereupon it picks up energy enough to become ionized and fragmented. This harsh ionization process often results in a number of fragments that may form a "fingerprint" suitable for the analyte identification.

Electron multiplier A device that transforms the ion beam of charged fragments into an electrical signal to be sent in a wire and monitored by a recorder or data system.

Forensic science Scientific methods that aim to solve legal questions.

General unknown A substance which perhaps is present in a sample but whose identity then is not known. Its revelation, which most often is an analytical challenge, may help explain a crime, death, or unexpected event.

Ion trap A type of mass filter. It has a closed space, in which an oscillating electric field is generated to direct the flow of ions made up by charged fragments, formed from an analyte and accelerated into the field. Only those ions with mass/charge ratios that do not suit the electric potentials of the oscillating field are thrown out of the cell to become recorded.

Magnetic sector The part of a mass filter that generates a magnetic field used to direct the flow of ions made up by charged fragments, formed from an analyte and accelerated into the field. Only those ions with mass/charge ratios that suit the acceleration voltage and magnetic field strength pass the magnetic sector to become recorded.

Mass/charge ratio The mass of a fragment divided by the number of charges it carries. In most instances only one charge is involved, and the mass/charge ratio is, therefore, generally equal with the mass of the fragment.

Quadrupole A type of mass filter. It has an open space between four rods where an oscillating electric field is generated to direct the flow of ions made up by charged fragments, formed from an analyte, and accelerated into the field. Only those ions with mass/charge ratios that suit the electric potentials of the oscillating field pass the rods to become recorded.

MASS SPECTROMETRY is a main tool in forensic chemistry for the analysis of chemicals in samples taken to reveal a possible crime. Its main value rests on the fact that mass spectrometry allows for an unbiased search with high sensitivity and specificity for a variety of substances with wide ranges of different physicochemical properties. The principles for making a substance suitable for mass spectrometric analysis are presented. These include the sample inlet from atmospheric pressure to the low pressure in the apparatus, the conversion of the sample molecules to ionized particles, their separation from one another in a magnetic or oscillating electric field, and their detection. In forensic chemistry the mass spectrometer is most often focused on a wide range of different mass fragments, and the data gotten from a test, therefore, need refinement to become intelligible. Examples from real-life forensic work in toxicology, arson analysis, and environmental forensics have been chosen to illustrate such processes.

Forensic science plays a key role for law-enforcing bodies. Its main task is to supply physical evidence pertaining to a suspected criminal act. The forensic scientist searches materials collected by the police for evidence a person may have left at the site of a crime or brought with him. The test matter is often made up of agents suitable for chemical analysis, such as drugs seized by the police or present in a body fluid, residues of arson accelerants or explosives, poisons in a dead body, or residues of illegally deposited spill oil, to name a few examples.

Two distinct approaches can be used for the chemical exam. In some surveys the analyst, without having to identify each single compound, may just compare the pattern of a group of substances in the sample with that of a reference product that has some bearing on a crime. In some types of suspected crimes (e.g., intoxication or drug trafficking), the individual substances in a sample must be pinpointed. Regardless of the analytical aim, a method must be used that will hold up to scrutiny in a court of law. Another problem the chemist is faced with is that he does not know which compounds to look for amid the wide range of possible ones with different physicochemical properties. By virtue of its high sensitivity and specificity, as well as its ability to reveal the general unknown, mass spectrometry (MS) is the chemist's primary analytical tool for solving many forensic problems.

Mass spectrometry is, in short, a method by which one generates charged molecules and molecular fragments and then measures the mass of each or, more rightly, their mass/charge ratio. The uncharged gas molecules, which are introduced into the mass spectrometer, move randomly around in space, so to control their motions they have to be ionized (i.e., a charge is added to the molecules). Also, energy in excess of the ionization must be supplied to break some of the covalent bonds holding a molecule together. This is to split it into specific fragments that may be used to obtain structural information about the molecule or to unfold its characteristic mass spectrometric fingerprint.

An outline of the process is shown in Fig. 1. Our discussion will include a general description of the hardware to generate the MS raw data (the introduction of the sample at atmospheric pressure into the mass spectrometer, which is under vacuum), the formation of the mass fragments and their separation from one another, focusing, and detection. The raw data thus obtained require some form of evaluation to become intelligible, a process that calls for automatic and/or manual computer work. A few examples of real-life forensic studies will help explain common approaches used in such a pursuit.

I. DATA GENERATION

A. Sample Inlet

The inlet system is used to introduce the sample into the mass spectrometer, to convert it into the gas phase, and to reduce its pressure before ionization. Forensic samples are often impure, so the analytes, have to be separated from the matrix before being inserted into the mass spectrometer. The inlet system is most often an interface between a chromatographic device and the mass spectrometer. By this approach, the analytes are separated from one another and from the contaminants by either gas chromatography (GC) or high-performance liquid chromatography (HPLC), and the isolated compounds in the effluents from the column flow directly into the mass spectrometer.

The combination of GC and MS (GC/MS) is often used in forensic science. Since the GC separation today is done in a capillary, and not in a packed column as it used to be, introducing the analytes into the mass spectrometer has become rather simplified. The flow rates are much lower in a capillary than in a packed column, and the entire volume of the effluents can be let into the mass spectrometer without any losses simply by pushing the loose end of the GC capillary in close to the ion source of the mass spectrometer.

Even though HPLC in combination with mass spectrometry (HPLC/MS) is used less frequently than GC/MS,

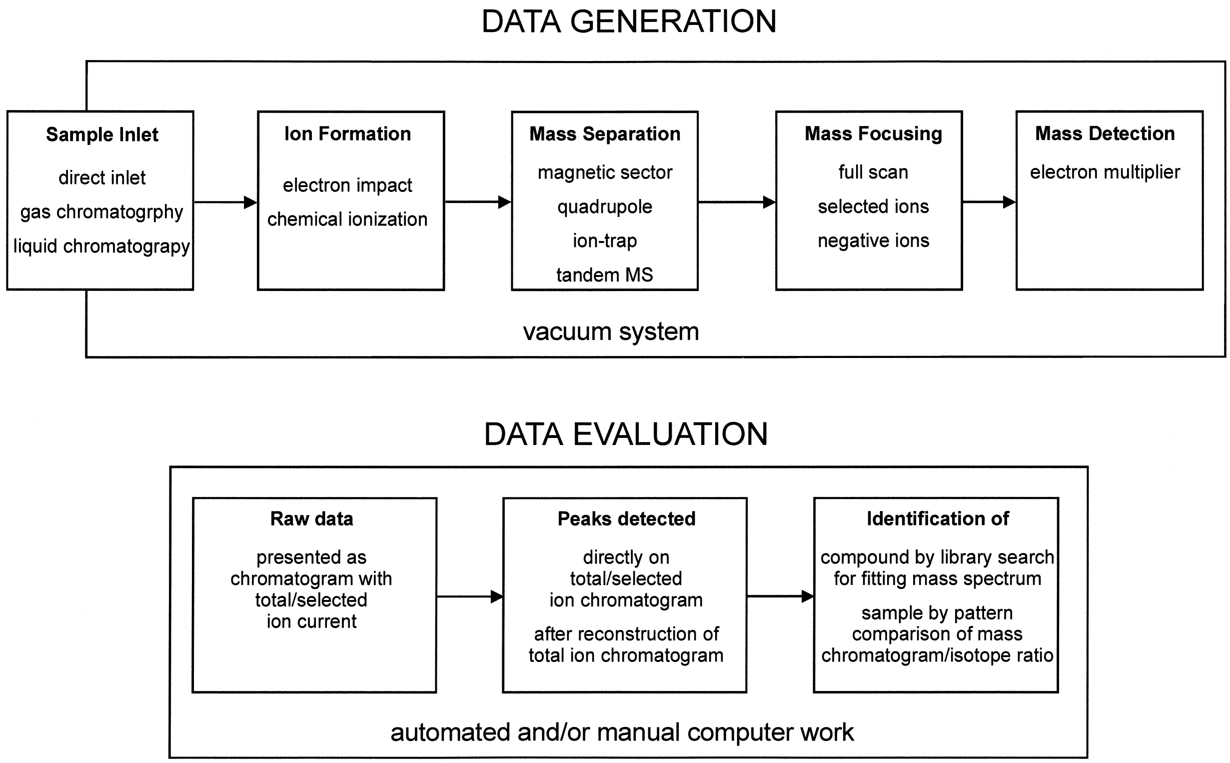


FIGURE 1 Scheme for common mass spectrometry approach in forensic chemistry.

it has recently gained in popularity. A reason for this delay in its use in the forensic science arena is the technical difficulties of interfacing a mass spectrometer with a liquid chromatograph. The eluate from the HPLC column is a liquid, which expands when it gasifies to become ready for MS analysis, a situation that adds an extra burden on the vacuum system for reducing the pressure in the mass spectrometer. Moreover, the effluents often carry polar, heat-labile substances that may taint the ion source when they are vaporized.

The thermo spray is one of the most often used HPLC/MS interfaces. In this system the HPLC effluents, upon entering the mass spectrometer, are forced through a pinhole leak to form a jet of liquids which is heated to gasify the solvents in the aerosol. The mist of droplets in the gas phase, which becomes void of solvents, carries, along with a charge from any ions initially present in the solution, the less volatile analytes directly into the ion source of the mass spectrometer to become fit for analysis.

B. Ion Generation

The mass fragments of a target substance are generated in the ion source of the mass spectrometer. An outline of the simplest ionization method, electron impact (EI), is shown in Fig. 2. This step puts a charge on the molecule and breaks some of the chemical bonds of the molecule by putting energy into it. In this process, the analyte molecule (M) is impinged with a beam of energetic electrons (e^-), a process that results in the formation of a radical cation with an odd number of electrons ($M^{+\cdot}$). As seen in the figure, the end result yields an analyte molecule that has lost an electron and become a particle with a positive charge, thus making it feasible for separation in a magnetic or oscillating electric field based on its mass and number of charges. The energy of the electrons is generally set at 70 eV, an optional value chosen because it is high enough to exceed the ionization energy at about 10 eV and to split the molecule by breaking its chemical bonds.

For the identification of an analyte it is often useful to know its molecular mass. This piece of information, however, is generally not gained by EI; the process often results in a complex fragmentation of the molecule, making the mass deduction of the master substance difficult. To overcome this problem a milder ionization method must be used, such as chemical ionization (CI). This process is done in an apparatus similar to that for EI, but unlike the EI device, which holds the same low pressure inside as outside the ionization cell, CI is run at a higher pressure within the chamber than outside it. Letting the analytes and electrons into the ionization chamber through small holes allows this.

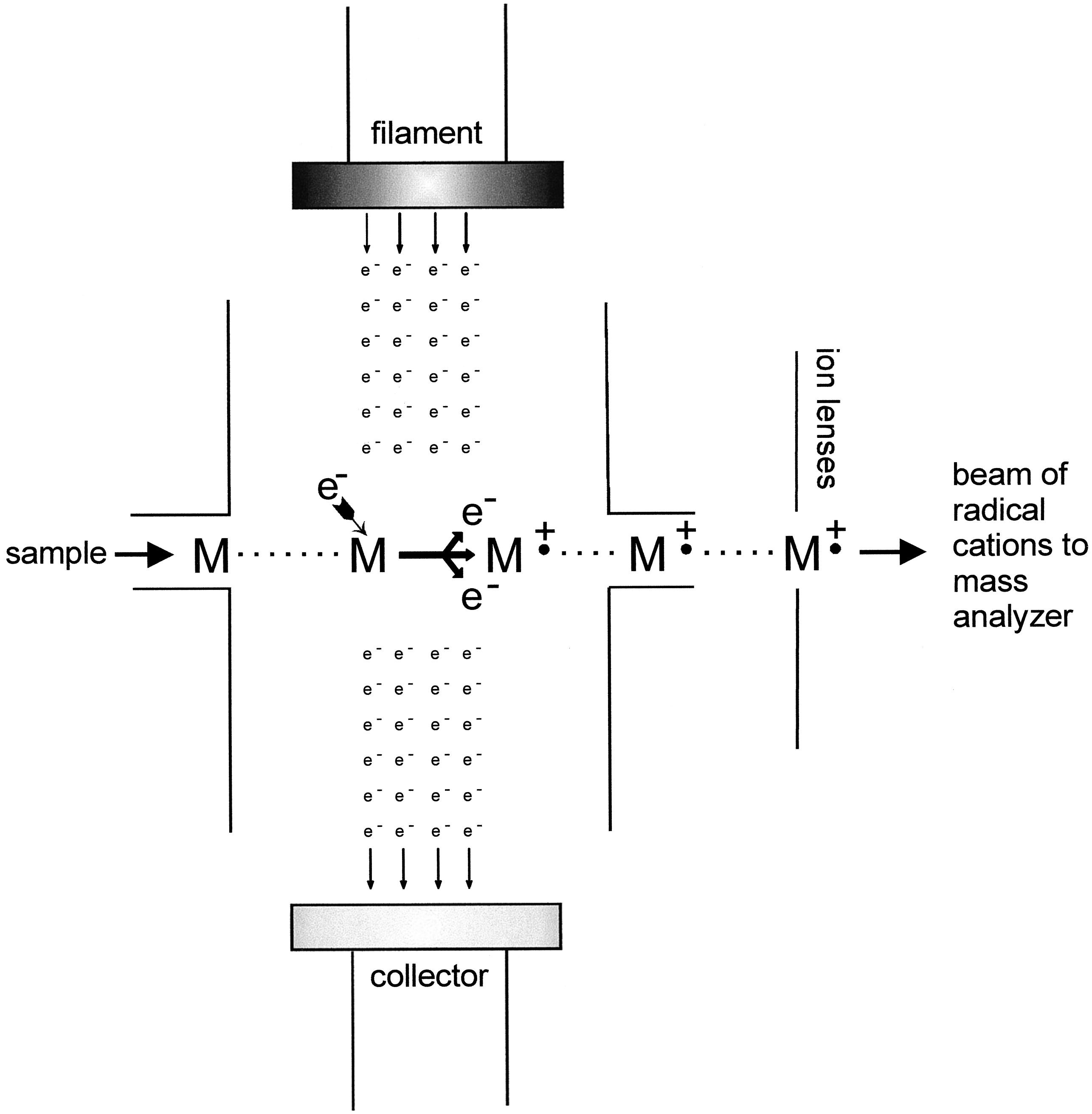


FIGURE 2 Fragmentation of molecules by electron impact. The analyte molecule (M) is bombarded with a beam of energetic electrons (e^-), whereby it loses an electron to become a positively charged particle ($M^{+\bullet}$), feasible for separation in a magnetic or oscillating electric field.

As shown schematically in Fig. 3 a proton donator formed by EI of the reagent gas (e.g., methane) collides in a gas phase with the target substance and then gives up its proton to the analyte. The reactions occur in two steps. In the first one, the primary ions are formed from the methane fragments. At a sufficiently high pressure in the ionization cell, the primary ions collide with neutral methane molecules to generate a stable population of secondary ions. These in turn react with the analyte gas, a process that results in a gas-phase acid–base proton transfer from the secondary ions to the analytes. Since these are present in low concentrations as compared with the concentration of methane, the analytes only seldom collide with the electrons, a situation that results in less

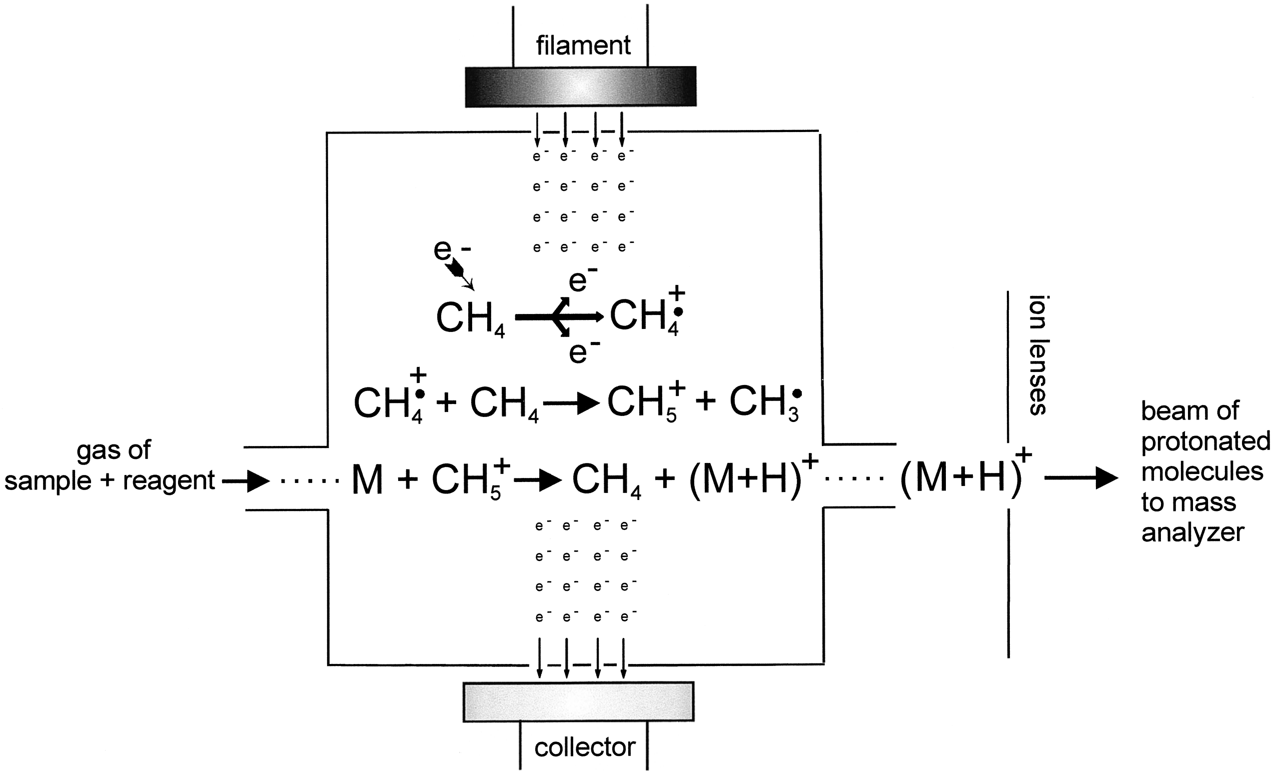


FIGURE 3 Fragmentation of molecules by chemical ionization. Gas of the analyte molecules (M) and reagent (CH_4) are let into the ionization chamber, where CH_4, bombarded with a beam of electrons, forms primary ions $CH_4^{+\bullet}$. At a sufficiently high pressure, these collide with CH_4 molecules to form the secondary ions (CH_5^+). These in turn collide with the analyte molecule (M), which then is converted to the protonated molecule $(M + H)^+$, suitable for separation in a magnetic or oscillating electric field.

fragmentation by CI than by EI. The protonated molecule, which appears as the most abundant fragment at CI, meets the demands for separation in a mass analyzer.

C. Mass Analyzer

The analyte molecules, which have become ionized, may now be moved apart from one another in a magnetic or oscillating electric field. Even though the object is to analyze the fragments based on their masses, it is the mass/charge (m/z) ratio that forms the basis for the separation process. This means that fragments with $m/z = 200/2$ and $m/z = 100/1$ have the same trajectories and may not be distinguished from each other. Fragments with two charges are, however, rare, and in practice it is the mass of a fragment that in general is regarded as the regulating factor of its movement in a magnetic or oscillating electric field. Appliances can have two fundamentally different operating MS modes—for example, mass spectrometers, that detect stable ions and those, that detect the unstable ones. The first type, or the beam-type scanning mass spectrometer, embraces the magnetic sector and the quadrupole instrument. The second type, or the mass-selective instability-operating machine, is known as the ion trap detector (ITD).

1. Magnetic Sector Instrument

The working principle for separating charged analyte fragments or molecules in the magnetic sector instrument is shown in Fig. 4. When thrown out from the ion source and accelerated, these tend to adjust to orbital movements when allowed to travel in a magnetic field aimed perpendicular to the flight of the ions. The trajectory of a fragment with a given m/z value, thus, depends on the acceleration voltage and the strength of the magnetic field. This means that, at a given acceleration, only those fragments with m/z values that balance the centripetal force of the magnetic field will travel through the flight tube and reach the detector, whereas the unstable ions will adhere to the path wall.

2. Quadrupole Instrument

Another type of mass filter is the quadrupole, whose working principle is shown in Fig. 5. Its ability to separate fragments with different masses is based on the fact that,

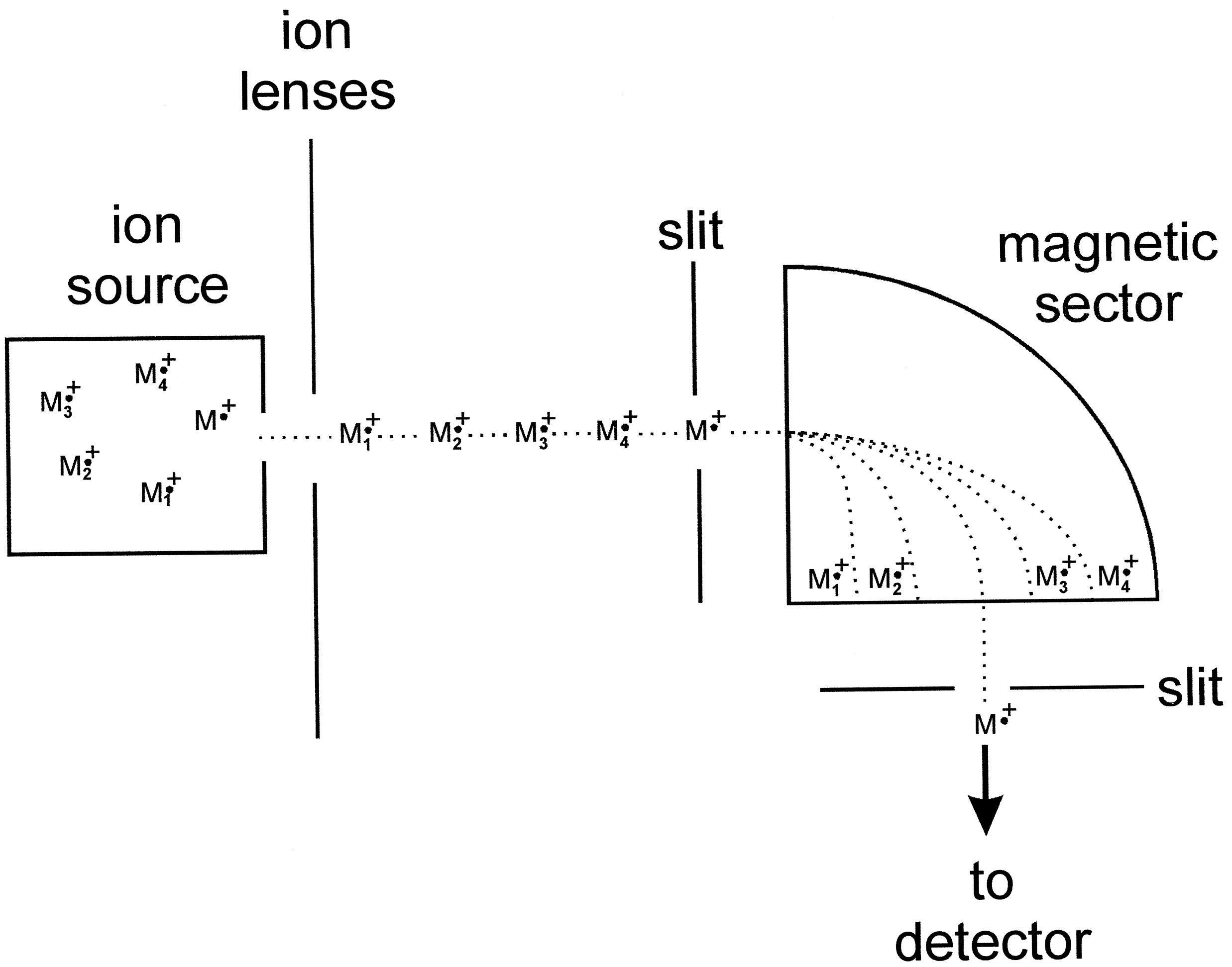


FIGURE 4 Separation of mass fragments with magnetic sector instrument. Ions ($M^{\bullet+}$, $M_1^{\bullet+}$, $M_2^{\bullet+}$, ...) are accelerated out of the ion source into the flight tube of the magnetic sector instrument, whose magnetic field, perpendicular to the trajectories of the ions, forces the charged particles to move in circular orbit. Only those ions ($M^{\bullet+}$) with a given m/z value will, at a given acceleration and magnetic strength, become stable and exit the magnetic mass separator to be detected, whereas the unstable ones ($M_1^{\bullet+}$, $M_2^{\bullet+}$, ...) will stick to the wall of the flight tube.

when ions travel in a beam of an oscillating electric field, their trajectories become bent. The quadrupole mass filter is made up by four parallel rods, about 10 cm long and 1 cm in diameter, which in a cross section are arranged to form a square box with a gap in between the four bars. The fragments are allowed to travel in this space along the rods, where they are accelerated in the oscillating electric field set up by dc and ac currents applied to the rods. At a certain ac and dc potential, ions with a specific m/z value will become stable and oscillate in a fixed path through the quadrupole to reach the detector to be recorded. The fragments with m/z values that do not suit the applied potentials will become unstable and, as in the magnetic sector instrument, will not reach the detector but will stick to the rods.

3. Ion Trap Detector

The working principle of a third type of mass filter, the ion trap detector (ITD), is illustrated in Fig. 6. It operates like the quadrupole based on the concept that the trajectories of ions, traveling in an oscillating electric field, become influenced by the wave frequency. Dissimilar to the quadrupole filter, however, the ion separation of the ITD occurs in a closed cavity, where the dc and ac currents applied to a ring electrode and an end cap of the cell set up the electric field. The molecules enter the cell in which they are ionized by an electron beam, and under the control of the given electric field the ions are forced to move in an orbit within the space of the cell. When the ac or dc potential is changed, the motion of some ions becomes unstable,

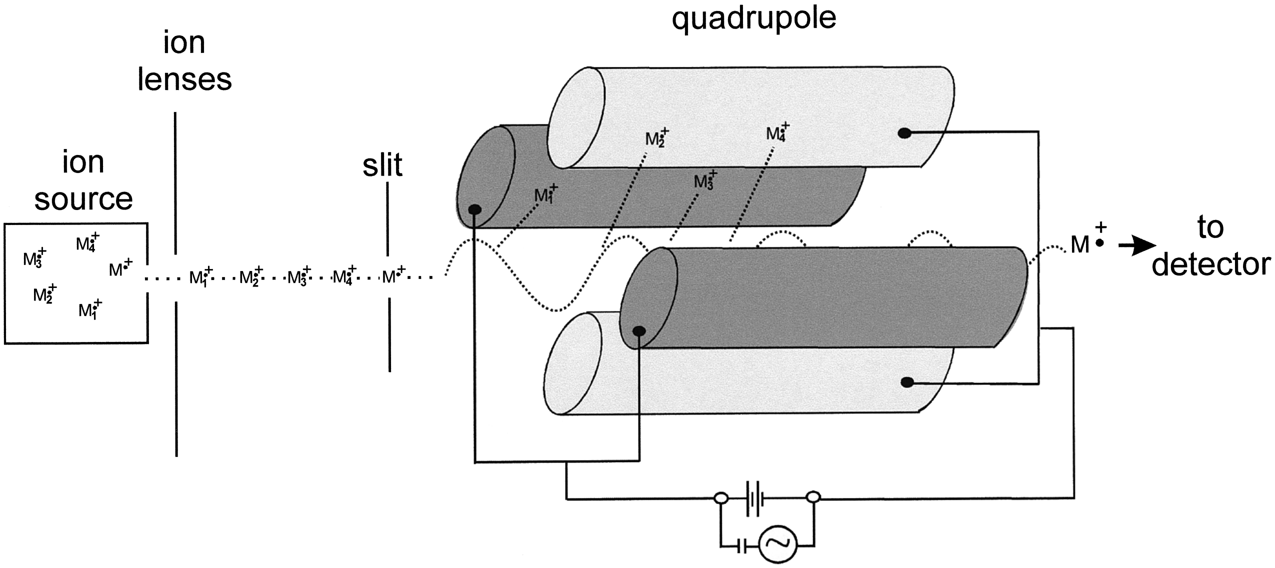


FIGURE 5 Separation of mass fragments with quadrupole instrument. Ions ($M^{+\bullet}$, $M_1^{+\bullet}$, $M_2^{+\bullet}$, ...) are accelerated out of the ion source into the space between four parallel rods with an oscillating electric field. At a certain ac and dc potential, ions with a specific m/z ($M^{+\bullet}$) value will become stable and oscillate in a fixed path through the quadrupole to reach the detector to be recorded, whereas the unstable ones ($M_1^{+\bullet}$, $M_2^{+\bullet}$, ...) will stick to the rods.

and 50% of these are then ejected from the cell through a hole in its bottom to be recorded by a detector. Unlike the quadrupole or magnetic sector instrument, in which the separation and detection of an ion occur in a continues process, the ITD operates in two steps separated in time: ion accumulation and mass analysis. The ITD also differs from the beam-type scanning MS by detecting the unstable ions, whereas the magnetic sector or quadruple instrument scans the stable ones, conditions that may offer rather different analytical possibilities.

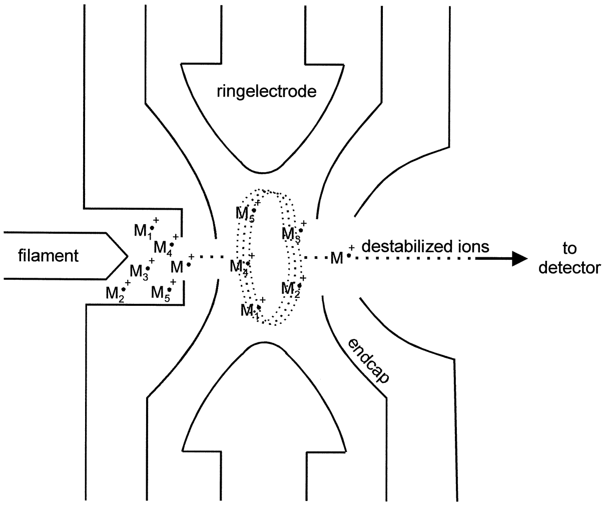


FIGURE 6 Separation of mass fragments with ion trap detector. Ions ($M^{+\bullet}$, $M_1^{+\bullet}$, $M_2^{+\bullet}$, ...) are introduced into a closed cavity and, under the control of the applied electric field, the ions are forced to move in an orbit within the space of the cell. When the ac or dc potential is changed, the motion of ions with given m/z values becomes unstable ($M^{+\bullet}$), and they are then ejected from the cell through a hole in its bottom to be recorded by a detector.

4. Tandem Mass Spectrometer

In tandem mass spectrometry (MS/MS), two mass spectrometers and a collision chamber are hooked up in series. Early devices utilized two magnetic sector instruments, but because these instruments became bulky and high priced, modern tandem devices are nearly all based on the use of quadrupole or ITD apparatus. The operating scheme of MS/MS is shown in Fig. 7. An advantage of MS/MS is that it provides high selectivity and extra information about the nature of an analyte, as well as reduced chemical background noise, which results in an increase in the signal-to-noise ratio for a detected peak. Even though this method, as judged from the number of scientific papers reported, does not seem to be in common use for forensics today, MS/MS will probably become the method of choice in the years to come.

D. Mass Focusing

The MS test can be run in the full scan mode or with selected ion monitoring (SIM). In the full scan mode, the mass analysis covers a range of m/z values, whereas one or a limited number of m/z values are selected for the exam with SIM. When picking between the two options,

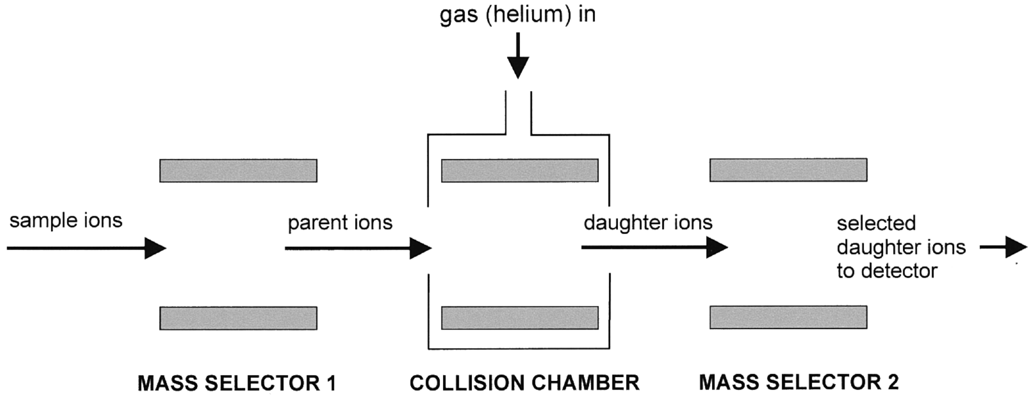


FIGURE 7 Tandem mass separator. Sample ions enter mass selector 1, where the parent ions are separated. The selected ions next enter the collision chamber, where they collide with gas molecules to form the daughter ions, which are finally separated in the mass selector 2 and expelled for detection. Either magnetic sector or quadrupole mass analyzers or both types mixed can make up a tandem mass spectrometer. A single ion trap can also function as a tandem mass spectrometer performing the same processes as described above in the same location but in consecutive steps.

operators of the magnetic sector instrument or the quadrupole have to consider to what extent they are willing to trade sensitivity for selectivity or vice versa and whether they are searching for the general unknown or for a suspected agent. The reason for this is that the window of the stable m/z values is sequentially swept across the entire m/z range of interest. The ratio of the transmitted window width to the width of the entire m/z range (i.e., the duty cycle) is in most scanning tests only a fraction of 1%. More than 99% of the ions from the target agents are lost. A duty cycle of 100% is possible with a beam-type instrument, but only when run in a nonscanning mode, as with SIM.

With the ITD, on the other hand, deciding whether to work in the full scan or SIM mode becomes less crucial. As described earlier, the ITD monitors the ions with unstable trajectories in two serial steps. Because these operating steps are separated in time, the yield of detectable ions will become high and rather independent on the scan range. Use of the ITD allows scanning with a high sensitivity the entire mass range that covers a substance group of interest. Given that forensic scientists often do not know what to look for and therefore need a search method with high sensitivity, the ITD should perhaps best meet such demands. A drawback of the ITD is that its sensitivity is more dependent on interfering substances than the beam-type scanning mass spectrometer; the ITD sensitivity thus drops with increasing amounts of impurities that may be co-eluted with the analytes during the chromatographic separation. The generation of somewhat distorted mass spectra at high analyte concentrations, giving rise to enhanced $[M+1]^+$ peaks, is another ITD problem.

In addition to the positive ions formed during the ionization of a molecule by EI or CI, negative ions are also produced and, by changing the electric field of the mass spectrometer, these can be monitored. At certain instances, such as when the target substances have a high affinity for electrons, negative ion monitoring can be extremely useful, mainly because of the high sensitivity that can be achieved. The approach of using negative ion monitoring has been particularly fruitful for the analysis of halogenated drug substances, which have been detected at 100- to 1000-fold higher sensitivity than when tested by the positive ion monitoring.

E. Mass Detection

In forensic science work, the electron multiplier is the most often used tool for detecting the separated ions, and a common device is the so-called horn-type electron multiplier; this variety of detector is the most compact and low cost. When the charged fragments enter the detector and strike the surface area of the horn, as shown in Fig. 8, electrons are emitted and accelerated by an electrical potential difference in the horn. These in turn hit the surface and new electrons are formed, a process that is repeated over and over again to generate a cascade of a progressively raised number of electrons, which finally are recorded as a signal. Usually the gain of the emitted electrons is on the order of 10^4 to 10^7 per ion entering the detector.

II. DATA EVALUATION

Mass spectrometry is used as an analytical tool in many forensic situations. In some instances, however, it can be regarded only as a complement to other chemical methods.

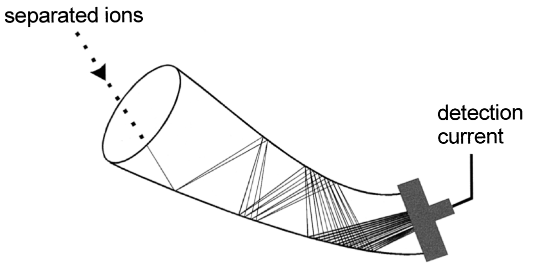


FIGURE 8 Detector: schematic figure of electron multiplier.

Three examples have been selected to show how MS has become an indispensable analytical tool.

A. Toxicology

The search for drug substances, pesticides, poisons, and their metabolites in body fluids from living persons or in postmortem organs presents an important and difficult task for the chemist. In addition to the fact that the forensic scientist most often does not know what intoxicant to look for, the main reason for using MS to begin with is the large number of possible toxic substances. In toxicology work at the Poison Center in Munich, as many as 8000 different substances have, in fact, been reported in 40,000 investigated objects.

Figure 9 shows an outline of the usual MS approach for searching a biological sample taken from a human for alien compounds with pharmacological effects. The example selected is a real-life incident of reckless driving by a motorist apprehended by the police on the suspicion of being under the influence of some drug or drugs. An extract of the blood sample was injected into a gas chromatograph/mass spectrometer focused on a broad range of mass fragments. The mass chromatogram at A in the figure is made up of the total ion current (the sum of all fragments recorded) and showed no clear peaks indicative of any drug substances. To raise the signal-to-noise ratio, the total ion current was reconstructed with the sum of the m/z 91 and 92, and then a peak appeared on the new mass chromatogram at B.

In the next step of the analytical process, the substance generating the peak at B was to be identified, which was achieved by comparing the mass spectrum of the analyte at C with mass spectra in an on-line library. Out of ten candidates picked by the program, three possible ones are shown at D, E, and F. Even though the mass spectrum of the analyte best fitted that of methylbenzene, it also matched nearly as well the mass spectra of the two other candidates. In addition to the recorded fragments at m/z 91 and 92, the final identification of the analyte was based on the fact that the retention time for the analyte was the same as that for methylbenzene. To hold up to legal scrutiny, proof of the analyte identity generally must indicate that at least two fragments and the retention time are the same as for the suggested substance. In the example here, the motorist suspected of being under the influence of drugs was probably a "sniffer," who had inhaled paint thinner or some other solvent containing toluene (methylbenzene) before driving his car.

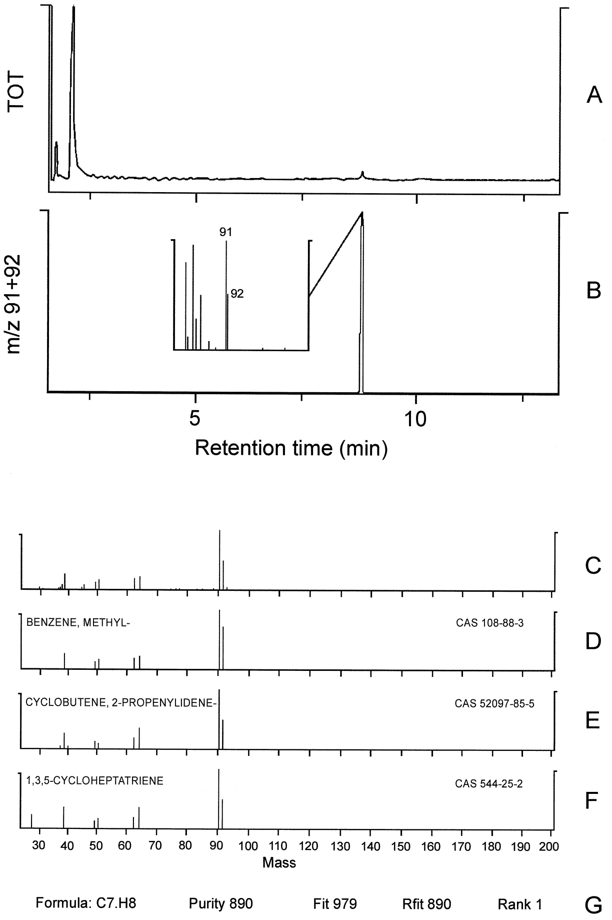


FIGURE 9 Search of blood sample for toxics. The mass chromatogram at A shows the total ion current (the sum of all fragments recorded) and the reconstructed mass chromatogram at B shows the ions with the sum of m/z 91 and 92. The mass spectra at C–F depict the library search for identifying the peak at B. The unknown analyte's mass spectrum is, after background subtraction, displayed at C. The three hottest candidates in the library along with their names and CAS (Chemical Abstracts Service) numbers are shown at D–F. At G are shown the chemical formula of the first ranked candidate and the value for how well the mass spectrum of the candidate fits with that of the analyte and vice versa. A value of 1000 indicates identical mass spectra; zero, no fragments in common.

B. Arson Analysis

The term *arson analysis* implies the search of materials taken from a fire scene for accelerant residues to establish

TABLE I Common Accelerants

Main type (approximative boiling point)	Major components	Examples of commercial products	Abundant mass fragments
Light petroleum distillates (<120°C)	*n*-Alkanes, branched alkanes	Petroleum ethers, pocket lighter, rubber cement solvents, lacquer thinners	42, 43, 56, 57, 71, 85
Gasolines (50–220°C)	Alkylbenzenes, naphthalenes	Automotive gasoline, lantern fuels	91, 92, 105, 106, 115, 116, 118, 119, 120, 134, 141, 142, 148, 156
Medium petroleum distillates (60–200°C)	*n*-Alkanes, branched alkanes, alkylbenzenes	Charcoal starters, paint thinners, mineral spirits, torch fuels, dry-cleaning solvents	43, 55, 57, 71, 82, 83, 91, 105, 120, 134, 138
Kerosene (90–290°C)	*n*-Alkanes, branched alkanes, alkylbenzenes, naphthalenes	Fuel oil, aviation fuel, insect sprays, charcoal starters	57, 71, 91, 106, 120, 128, 134, 142, 148
Heavy petroleum distillates (120–410°C)	*n*-Alkanes, branched alkanes, naphthalenes	Fuel oil, diesel fuel	43, 57, 71, 128, 142
Varia	Alcohols, ethers, α-pinene	Solvents, turpentine	31, 32, 45, 59, 73, 74, 93

whether the incident was of incendiary origin or not. The U.S. National Fire Protection Association reported that over 100,000 fires in 1994 were arson related and more than 500 persons lost their lives in these fires. In terms of loss of human life, houses, properties, and goods, this type of incidence, extracts a large toll from society as well as from individuals, and makes the forensic task urgent.

The list of the most common accelerants, shown in Table I, may at first give the impression that these would rapidly gasify along with the burning solid materials and thus not become detectable, but it is not always so. Traces of the fuel often remain in some closed areas of the fire scene even after the temperature has reached perhaps 1000°C, and these accelerant residues may be detectable after proper sample collection.

By virtue of its high sensitivity and specificity, GC/MS is well suited for searching trace amounts of test materials for residues of accelerant components. As reported in the literature on material spiked with accelerants, specific GC/MS patterns of different types of fuels may be used to identify an accelerant. A problem with this approach is that burning plastics and other solid materials release hydrocarbons that are the same as those in accelerants and thus may contribute to a false positive result by tainting the test material. The use of fuel labels may overcome this hurdle. Methyl-tertbutyl-ether (MTBE), which is an additive in gasoline, has been suggested as such a marker. Owing to its low boiling point (55.2°C), however, it disappears rapidly during a fire and is therefore only detectable in postmortem materials from a fire victim who has inhaled MTBE. Figure 10 shows such an example from the real-life fire of a villa, which was completely destroyed. No accelerant residues were spotted in the ashes around the body of a fire victim, so blood samples of the deceased were also searched for fuel components. On the chromatogram monitoring the sum of all ions (TOT), no peaks appeared, but after it had been reconstructed with the selected ions at m/z $78+91+92$ a number of aromatic hydrocarbons showed up. These could, however, stem either from a fuel or from pyrolyzed plastics. The presence of MTBE at m/z 73 indicated, however, that gasoline had been used to set the fire. This piece of evidence pointing to arson could not have been brought light without MS.

C. Environmental Issue

Growing global use and transportation of chemicals with potential deleterious effects on the environment have prompted forensic scientists to develop methods for

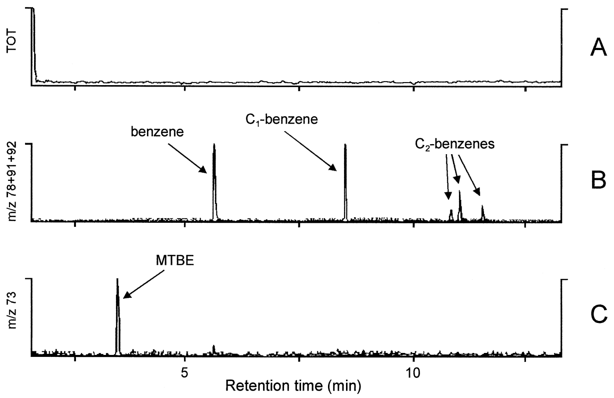


FIGURE 10 Arson analysis of blood sample. The mass chromatogram at A shows the total ion current (the sum of all fragments recorded). The reconstructed mass chromatogram at B shows the ions with the sum at m/z 78, 91, and 92 to monitor aromatic hydrocarbons, and the reconstructed mass chromatogram at C shows the ions at m/z 73 to monitor MTBE.

detecting the origin of illicit waste discharge. A common issue is to link the constituents of, for instance, an oil spill sample with its original source. GC or GC/MS are usually used to do the analysis when the individual hydrocarbons are separated and displayed on a chromatogram, whose pattern is compared with that of the suspected original source. A major problem in such a survey, however, is that some time usually has passed between the waste discharge and the specimen collection. Owing to weathering by evaporation and biodegradation of the components during this period, the composition of the oil spill may have changed to the extent that it bears only limited similarities with that of the origin.

Monitoring the contents of stable isotopes by MS in the whole sample or in its individual constituents after GC has isolated them is an approach used to get around the weathering problems. It has been shown that the most common stable isotope parameter used (i.e., the $^{13}C/^{12}C$ ratio) may be specific for a source and not very much influenced by weathering effects. Technically, the analysis is done after the test material has been combusted to carbon dioxide and water, and the $^{13}C/^{12}C$ ratio is then determined in the carbon dioxide fraction by MS. To determine the $^{13}C/^{12}C$ ratio of the individual components of a sample, they are separated by GC, each isolated hydrocarbon goes directly into a combustion chamber, and the carbon dioxide, after having been freed from water, enters the mass spectrometer. An illustration of the use of this tool to identify oil spill sources is the stable isotopes of samples taken from the area of Prince William Sound several years after the Exxon *Valdez* disaster. The isotope distribution in the samples as measured on the bulk contents correlated with two distinct origins of the oil spill: one coming from the Exxon *Valdez* oil and the other from Californian crude oil. The latter source had probably been released in to the sea from storage tanks by an earthquake that had taken place in the area 40 years earlier.

SEE ALSO THE FOLLOWING ARTICLES

DNA Testing in Forensic Science • Gas Chromatography • Ion Kinetics and Energetics • Liquid Chromatography • Mass Spectrometry • Pollution, Environmental • Spectroscopy in Forensic Science • Toxicology in Forensic Science

BIBLIOGRAPHY

Currell, G. (2000). “Mass spectrometry systems.” *In* “Analytical Instrumentation, Performance, Characteristics and Quality,” pp. 181–191, John Wiley & Sons, New York.

Fifield, F. W., and Kealey, D. (2000). “Mass spectrometry.” *In* “Principles and Practice of Analytical Chemistry,” 5th ed., pp. 426–440, Blackwell Scientific, Oxford.

Keto, R. O. (1995). “GC/MS data interpretation for petroleum distillate identification in contaminated arson debris,” *Journal of Forensic Sciences* **40,** 412–423.

Masucci, J. A., and Caldwell, G. W. (1995). “Techniques for gas chromatography/mass spectrometry.” *In* “Modern Practice of Gas Chromatography,” 3rd ed. (R. L. Grob, ed.), pp. 323–391, John Wiley & Sons, New York.

McNair, H. M., and Miller, J. M. (1998). “Special topics.” *In* “Basic Gas Chromatography. Techniques in Analytical Chemistry,” pp. 153–163, John Wiley & Sons, New York.

Wasels, R., and Belleville, F. (1994). “Gas chromatographic–mass spectrometric procedures used for the identification and determination of morphine, codeine and 6-monoacetylmorphine,” *Journal of Chromatography A* **674,** 225–234.

Microwave Molecular Spectroscopy

Robert L. Cook
Mississippi State University

I. Background
II. Experimental Techniques
III. Evaluation of the Moments of Inertia
IV. Rigid-Rotor Energy Levels and Spectra
V. Centrifugal Distortion Effects
VI. Rotation–Vibration Interactions
VII. Internal Rotation
VIII. Evaluation of Molecular Structures
IX. Stark Effect and Zeeman Effect
X. Nuclear Quadrupole Hyperfine Structure
XI. Advanced Experimental Methods

GLOSSARY

Asymmetric top Molecule for which all three principal moments of inertia are different.

Asymmetry parameter Measure of the departure of a molecule from the rotational behavior of a symmetric-top molecule.

Centrifugal distortion Molecular rotation gives rise to a centrifugal force causing the bond lengths and bond angles to change slightly. This leads to shifts of the rotational lines.

Double resonance Technique in which two radiation fields (a pump and a probe) are introduced into the sample cell. Can be used to simplify complicated spectra.

Effective constants Values of molecular parameters averaged over the vibrational motion, and hence associated with a given vibrational state. For example, the effective rotation and distortion constants.

Equilibrium constants Values of molecular parameters associated with the minimum in the vibrational potential function, the so-called vibrationless state. For example, the equilibrium bond distance.

Hyperfine structure Splitting of spectral lines into two or more components. This can arise, for example, from the application of external fields (electric and

magnetic), the coupling of internal and overall rotation, or the effects of nuclear coupling.

Internal rotation Rotation of two parts of a molecule about a single bond. Such an internal rotation is often hindered by a barrier.

Klystron, backward wave oscillator, YIG oscillator Sources of tunable monochromatic electromagnetic radiation in the microwave region.

Microwave region Part of the electromagnetic region extending from about 1000 MHz ($\lambda = 30$ cm) to 1 million MHz ($\lambda = 0.3$ mm).

Microwave spectrometer A microwave source, absorption cell, and detector. Used to investigate the rotational spectrum.

Microwave spectrum Usually an absorption spectrum (i.e., the abstraction of energy from the radiation field). The spectrum is associated with transitions between rotational energy levels in a given vibrational and electronic state and consists of a series of lines, each characterized by a frequency and an intensity.

Rigid rotor Idealization in which a molecule is treated as a rigid, nonvibrating rotor.

Rotational constants Constants that mainly determine the rotational energy of a molecule. They are inversely proportional to the moments of inertia, which in turn depend on the mass and geometry of the molecule.

Rotational energy Quantum mechanics requires molecules to rotate only at certain rates and hence to have only discrete values of rotational energy and total angular momentum.

Rotational quantum numbers Numbers that specify various angular momenta and distinguish the rotational energy levels.

Satellite spectrum Rotational absorption lines that arise from an excited vibrational state rather than the ground vibrational state.

Selection rules Rules that specify whether or not a given transition between two particular energy levels is allowed.

Symmetric top Molecule for which two of the three principal moments of inertia are equal. Symmetric tops may be further divided into prolate and oblate tops.

Waveguide Rectangular-shaped metal pipe used to transmit microwave radiation. Different frequency regions require different-size waveguides.

MICROWAVE molecular spectroscopy involves the observation and analysis of absorption transitions between molecular rotational energy levels of gas-phase molecules. These transitions between rotational levels are associated with a given vibrational state of the ground electronic state, and the transitions fall in the microwave region of the electromagnetic spectrum, which lies between the conventional radiowave region and the infrared region. The distinguishing features that characterize microwave spectroscopy are high resolution and highly accurate frequency measurements. These characteristics, coupled with a sensitivity to molecular structure, isotopic composition, and various other molecular properties, make it a powerful technique for obtaining detailed molecular information.

I. BACKGROUND

In microwave molecular spectroscopy, absorption spectra arise from molecular rotation and correspond to transitions between the rotational energy levels associated with a given vibrational state of a particular electronic state (see Fig. 1). The rotational transitions, which fall in the microwave region, are induced through the interaction of the molecular electric dipole with the electric vector of the radiation. The microwave region extends roughly from 1000 MHz ($\lambda = 30$ cm) to 1 million MHz ($\lambda = 0.3$ mm). In this spectral region, frequencies are expressed in megahertz (MHz) or gigahertz (GHz) units, where the Hertz unit denotes cycles per second; 1 MHz = 10^6 Hz and 1 GHz = 10^9 Hz. In terms of wavelengths, the region may be conveniently divided into the centimeter-wave ($\lambda = 1$–30 cm), millimeter-wave ($\lambda = 1$–10 mm), and submillimeter-wave regions ($\lambda < 1$ mm). Note that 30 and 300 GHz correspond to 1 cm and 1 mm, respectively.

The rotational absorption spectrum depends on the principal moments of inertia and, hence, is characteristic of the

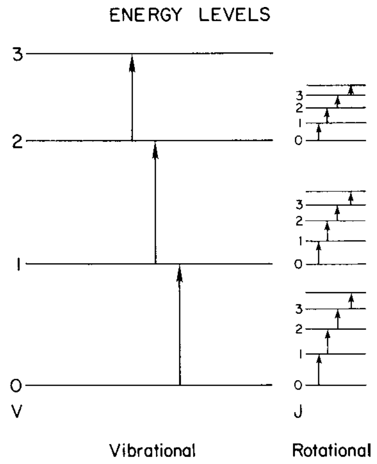


FIGURE 1 Vibrational energy levels associated with the ground electronic state. With each vibrational level there is a set of rotational energy levels. Pure rotational transitions, transitions between rotational levels of a given vibrational level, are on the right and pure vibrational transitions are on the left.

whole molecule and very sensitive to the molecular structure, isotopic composition, and numerous other molecular properties. Information that can be obtained from the observation and analysis of rotational transitions includes precise molecular structures, dipole moments, centrifugal distortion constants, vibrational potential functions, internal rotation barriers, nuclear masses and spins, nuclear quadrupole coupling constants, molecular magnetic moments, conformations of rotational isomers and ring compounds, magnetic susceptibility and electric polarizability constants, molecular quadrupole moments, structures and other properties of hydrogen-bonded complexes, rare atom–molecule complexes, molecular ions, nonpolar spherical-top molecules, and qualitative and quantitative analysis. The field of microwave spectroscopy continues to expand with new experimental and theoretical developments. The contributions of microwave molecular spectra to the fields of microwave–optical and microwave–infrared double-resonance spectroscopy, microwave molecular astronomy, and structures of weak complexes attest to its vitality.

II. EXPERIMENTAL TECHNIQUES

The microwave region has its own methods of generating, detecting, and measuring microwave radiation. During World War II considerable work was done to develop microwave radar systems. This led to new sources—the klystron and other hardware associated with the propagation and measurement of microwaves. After the war, this provided the impetus for a rapid development of microwave spectroscopy as an area of study.

A. Conventional Microwave Spectrometer

A typical microwave spectrometer is illustrated in Fig. 2. The essential elements of a microwave spectrometer are a microwave source, absorption cell, detection system, and a system for measuring the source frequency. Microwave sources—the klystron and, more recently, the backward-wave oscillator (BWO)—generate a very narrow band of frequencies so that the source is essentially monochromatic. Furthermore, the source frequency can be conveniently varied and is often phase stabilized to give good frequency stability.

Microwave radiation is conveniently transmitted through hollow, rectangular metallic conductors called waveguides. These are usually made of brass or copper, and the rectangular dimensions depend on the frequency range to be transmitted, the size decreasing with increasing frequency. Typical frequency bands used in the commonly studied low-frequency region are X-band (8–12.4 GHz), P-band (12.4–18 GHz), K-band (18–26.5 GHz), and R-band (26.5–40 GHz). The simplest absorption cell consists of an X-band waveguide 2–5 m long with mica windows at each end to provide a vacuum seal. Small holes can be made in the waveguide, usually on its wide face, to allow evacuation of the cell and for sample introduction. This cell is used for the various higher frequency bands mentioned previously by employment of tapered transition sections to match the cell to the smaller waveguide of the source and detector. Sometimes, the interior of the absorption cell is gold-plated to limit chemical decomposition.

The microwave radiation passing through the absorption cell is detected by a crystal detector mounted at the end of the waveguide. The small output voltage is amplified and displayed on an oscilloscope or chart recorder. A microwave spectrum is obtained by sweeping the microwave source over a range of frequencies and observing the small variations in power at the detector. An absorption line appears as a sharp dip on the recording system.

To increase the sensitivity of a microwave spectrometer, Stark modulation may be added. For this purpose, a thin metal septum is mounted in the absorption cell parallel to the broad side of the waveguide, using insulating material such as Teflon (see Fig. 2). An alternating (5–100 kHz) voltage, usually a zero-based square wave, is applied to the septum. Typically, voltages from a few to 2000 V are applied. As discussed later, in the presence of an external electric field, the rotational energy levels and, therefore, the absorption lines are split into a number of components. If, for example, the source frequency is tuned to the resonant frequency of an absorption line, the Stark voltage periodically shifts the absorption frequency away from the frequency of the source, resulting in a modulation (field on and off) of the absorption line at the frequency of the Stark modulator. In general, a signal reaches the amplifier when the source being swept reaches the field-off absorption frequency or the frequency of the field-on absorption lines (Stark components or lobes). The modulated signal is detected and amplified by a preamplifier tuned to the modulation frequency. By employing phase-sensitive detection, one can obtain additional sensitivity. Here, a final amplifier is referenced to the modulation frequency, and only the noise that has the same phase and frequency as the signal is amplified. With phase-sensitive detection, the Stark components are displayed in opposite phase to the zero-field line. Figure 3 shows the appearance of a rotational transition obtained with a Stark-modulated spectrometer with application of different modulation voltages.

Frequency measurements are made by comparing the frequency of the microwave source with the appropriate harmonic of a stable oscillator that is calibrated with a frequency standard. The difference frequency between the

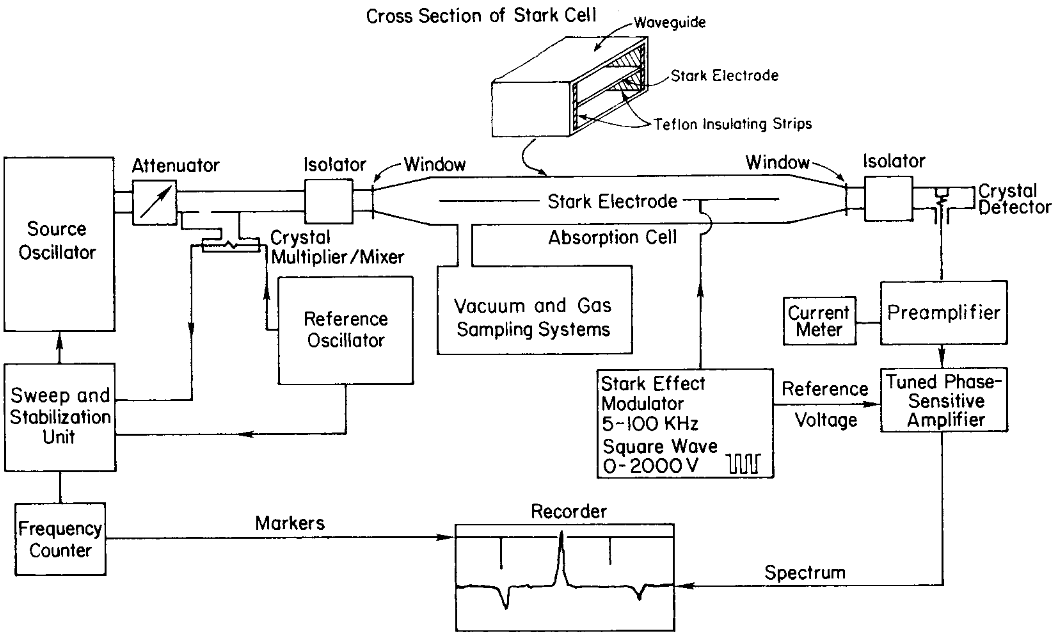


FIGURE 2 Basic elements of a Stark-modulated microwave spectrometer. The mounting of the Stark electrode in the absorption cell is shown in the inset.

harmonic signal and the source may be used as the input to a digital electronic counter. Frequency markers are often displayed on a chart recorder (see Fig. 2) along with the absorption line, and these are employed in the frequency measurements. Approximate frequency measurements of the source frequency can be readily made with a cavity wavemeter. Absorption lines in the microwave region are commonly measured with frequency accuracies of 1 part in 10^6 or better and with resolutions on the order of 0.1 MHz. Actually, higher resolution is obtainable with special spectrometers (see Section XI.D).

B. Millimeter/Submillimeter-Wave Spectrometer

As the frequency of the microwave radiation increases, the sensitivity of the technique increases rapidly [see Eq. (1)].

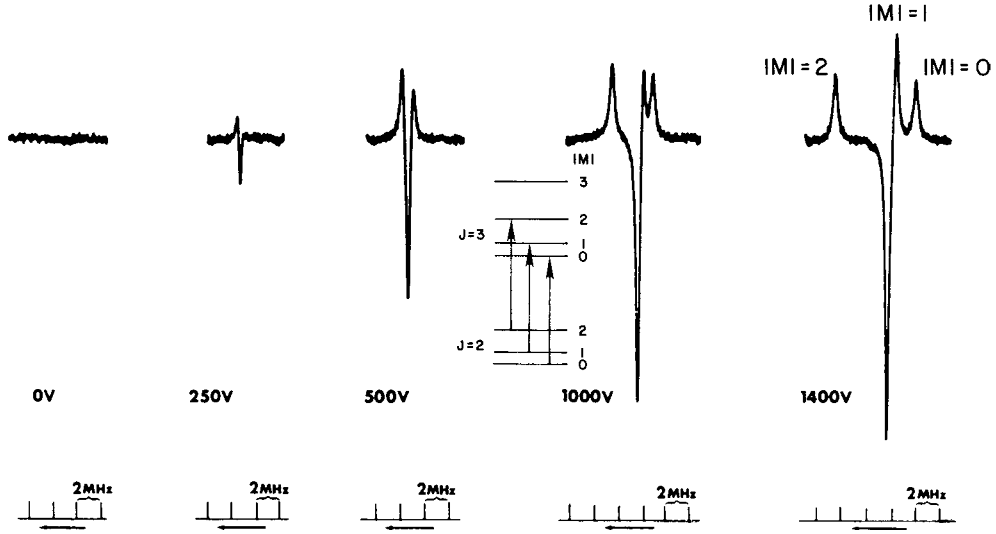


FIGURE 3 Effect of the square-wave Stark-modulation voltage on the appearance of the $J = 2 \rightarrow 3$ transition of OCS. Note that as the Stark lobes labeled by $|M|$ are displaced from the field-off line, the intensity of the line increases. The Stark effect is second order. The energy level diagram for a linear molecule in an electric field is shown in the inset. The allowed $\Delta M = 0$ transitions are also depicted.

Moreover, many small, fundamental molecules have spectra which fall in the shorter wavelength region. For these reasons, spectrometers have been developed operating at higher frequencies. This is accomplished by harmonic generation of the higher frequencies from a high-powered, lower frequency source using a point-contact crystal diode harmonic generator. Development of this and various other techniques to exploit the millimeter- and submillimeter-wave regions was pioneered by Walter Gordy and his research group. In the shorter wavelength region, quasi-optical techniques can be used, and more versatile absorption cell designs are possible. The Stark cell is usually replaced by a quasi-free space cell. Both metal and glass cells have been employed. Radiation transmission through the adsorption cell is properly matched to the waveguide by employing waveguide horns equipped with a Teflon lens (see Fig. 26). A block diagram of a broadband millimeter/submillimeter-wave spectrometer is given in Fig. 4. The YIG microwave oscillator is phase locked to the frequency synthesizer, and a computer is used to sweep the synthesizer. The output of the YIG is tripled and amplified by a TWT amplifier to ca. 1 W. The output is then further multiplied by a harmonic generator. The harmonics of the driving signal in the millimeter- and submillimeter-wave regions are propagated quasi-optically through the 1-m-long absorption cell. The spectrum is detected with lock-in detection and digitized and displayed by the computer.

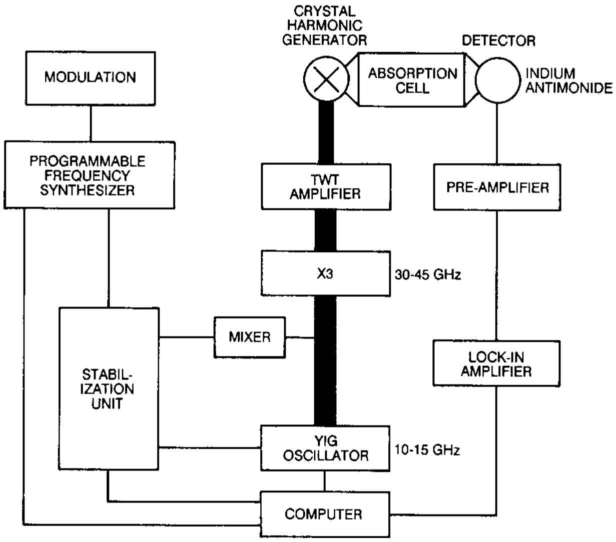


FIGURE 4 Schematic diagram of a computer-controlled broadband millimeter/submillimeter-wave spectrometer ($\nu_{max} \cong$ 440 GHz). The YIG oscillator is phase locked and swept by computer. A helium-cooled indium antimonide detector is used for signal detection, and the YIG is frequency modulated for improved signal recovery by lock-in detection. [After Booker, R. A., Crownover, R. L., De Lucia, F. C., and Helminger, P. (1988). *J. Mol. Spectrosc.* **128,** 62.]

C. Line Intensities and Shapes

A few comments on the shapes and factors affecting line intensities are in order. An absorption line is not perfectly sharp but is usually symmetrical about the resonant frequency. It is characterized by its resonant frequency ν_0, its shape, and its linewidth $2(\Delta\nu)$, where $\Delta\nu$ is the half-width at half-intensity points of the line. At very low pressures, <10 mTorr (1 mTorr $= 10^{-3}$ Torr $= 1\ \mu$m $= 10^{-3}$ mm of Hg), the linewidth arises primarily from Doppler broadening and is independent of pressure, and the line shape is Gaussian. This is the region of maximum resolution. As the pressure is raised in this region, the peak line intensity increases, while the linewidth remains constant. As the pressure is further increased, collision broadening becomes important. At pressures >10 mTorr, broadening due to molecular collisions becomes dominant. The line shape is now Lorentzian, the half-width is directly proportional to the pressure $\Delta\nu = kp$, and the peak intensity is independent of pressure. This is the region of maximum sensitivity.

A useful quantitative measure of the intensity of a rotational transition is given by the peak absorption coefficient α_0, defined by

$$\alpha_0 = \left[\frac{8\pi F_{j,\tau}\mu_g^2\,\lambda_g(J,\tau;J',\tau')}{3ck^2T^2(2J+1)}\right]\cdot\frac{\nu_0^2 x}{\Delta\nu/p}. \qquad (1)$$

This applies for any class of rotor with $F_{J,\tau}$, the fraction of molecules in the lower state J, τ of the transition and the vibrational state v; μ_g is the electric dipole moment component giving rise to the particular transition under observation, c is the speed of light; k is the Boltzmann constant; T is the absolute temperature of the gas, ν_0 is the frequency for which the absorption is a maximum; x is the mole fraction of the absorbing molecular species; p is the total pressure in the absortion cell; and $\Delta\nu$ is the half-width of the line. The above expression summarizes the various factors that affect the line intensity. Since $\Delta\nu$ is proportional to p, α_0 is independent of total pressure. Furthermore, the intensity increases with frequency, dipole moment, and λ_g $(J,\tau;J',\tau')$, the line strength of the transition. This latter quantity is related to the transition moment by

$$|\mu_{ij}|^2 = \mu_g^2\lambda_g(i;j)\big/(2J+1). \qquad (2)$$

The line strength for asymmetric rotors has been tabulated for various values of κ. For a symmetric top a simple relation applies:

$$\lambda(J,K;J+1,K) = [(J+1)^2 - K^2]/(J+1). \qquad (3)$$

This also applies to a linear molecule with $K = 0$. The intensity also depends on the fraction of molecules $F_{J,\tau}$, in the lower rotational state J, τ of the vibrational

state v. This fractional population is proportional to the Boltzmann factor $e^{-E/kT}$, where E is the sum of the rotational energy $E_{J,\tau}$ and the vibrational energy E_v. In general, lowering the temperature increases the ground vibrational state population and hence α_0, while increasing the temperature increases α_0 for transitions in excited vibrational states. Also, large molecules tend to have weaker spectra since $F_{J,\tau}$ is decreased. Typical values of α_0 in the centimenter-wave region are 10^{-5}–10^{-7} cm^{-1} or smaller. A weak line with a peak absorption coefficient as small as 10^{-9} cm^{-1} can, however, be detected with a good Stark-modulated spectrometer. To be specific, the isotopic species $^{18}O^{13}C^{32}S$ in natural abundance has an $\alpha_0 = 3 \times 10^{-9}$ cm^{-1} and, hence, should be observable with a signal-to-noise ratio of better than 3:1.

Space does not permit a discussion of the effects of nuclear spin on the intensity of rotational lines; however, a few comments are in order. For molecules with symmetry, identical nuclei are effectively interchanged as a result of the rotational motion. Since the total wavefunction must possess certain symmetry with respect to exchange of identical nuclei, this results in the rotational energy levels having different nuclear statistical weights. This, in turn, affects the relative intensities of the rotational transitions. Such effects are often used to confirm a particular molecular symmetry for large molecules. In the case of SO_2 (e.g., where the nuclear spin I of oxygen is zero) half of the allowed transitions are missing because of nuclear spin effects.

The above expression for α_0 is the basis for applications of microwave spectroscopy to chemical analysis. In qualitative analysis, the frequencies of the rotational absorption lines provide the basis for identification of the compound. In quantitative analysis, the intensities of the absorption lines provide the basis for determining concentration of the compound. Briefly, the term in brackets in Eq. (1) is constant if one compares the intensity for the same transition at the same temperature in both a standard and unknown sample. Therefore, one can write for the ratio of the mole fractions

$$x_1/x_2 = (\alpha_0)_1/(\alpha_0)_2 \cdot (\Delta\nu/p_1)/(\Delta\nu/p_2), \tag{4}$$

where $\Delta\nu/p$ is measured for each sample. The term $(\alpha_0)_1/(\alpha_0)_2$ is given by the ratio of the peak line heights from a Stark spectrometer when the microwave power level is held constant. The latter may be accomplished by keeping the crystal current constant and avoiding power saturation. From the known mole fraction of the reference sample it is possible by making the above measurements to evaluate the mole fraction of the unknown sample.

III. EVALUATION OF THE MOMENTS OF INERTIA

The first step in the study of a rotational spectrum is to evaluate the moments of inertia, or rotational constants, from which the rigid rotor spectrum (discussed in the following section) can be predicted. The rotational problem is treated mathematically in terms of a molecule-fixed axis system with its origin at the center of mass of the molecule and its axes oriented along the principal inertial axes. With respect to these axes the moments of inertia are constant, and the intertia matrix is diagonal. The principal axes of inertia are designated by a, b, and c. The corresponding moments of inertia are denoted by I_a, I_b, and I_c, where, by convention, the inertial axes are labeled so that $I_a \leqq I_b \leqq I_c$. In terms of the coordinates of the atoms in the principal axis system, the principal moments of inertia are defined by

$$\begin{aligned} I_a &= \sum m_i\left(b_i^2 + c_i^2\right),\\ I_b &= \sum m_i\left(a_i^2 + c_i^2\right),\\ I_c &= \sum m_i\left(a_i^2 + b_i^2\right), \end{aligned} \tag{5}$$

where a_i, b_i, and c_i are the coordinates of the ith atom of mass m_i, and the sum is over all atoms of the molecule. The principal axis system is also characterized by the auxiliary relations: the center-of-mass or first-moment equation

$$\sum m_i a_i = \sum m_i b_i = \sum m_i c_i = 0, \tag{6}$$

which insures that the origin is at the center of mass, and the product of inertia relations

$$\begin{aligned} I_{ab} &= -\sum m_i a_i b_i = 0,\\ I_{ac} &= -\sum m_i a_i c_i = 0,\\ I_{bc} &= -\sum m_i b_i c_i = 0, \end{aligned} \tag{7}$$

which insures that the inertia matrix is diagonal. In addition to the moment-of-inertia relations, the above equations are useful in the evaluation of molecular structures. The rotational energies, as we shall see, depend on the rotational constants designated by A, B, and C, with $A > B > C$, and defined by

$$A = \frac{h}{8\pi^2 I_a}, \qquad B = \frac{h}{8\pi^2 I_b}, \qquad C = \frac{h}{8\pi^2 I_c}. \tag{8}$$

This definition gives the rotational constants in frequency units, and the relation between A and I_a is

$$A\ (\text{MHz}) = 505{,}376/I_a\ (\text{amu Å}^2), \tag{9}$$

with similar expressions for B and C. This conversion factor is based on the ^{12}C mass scale.

The different types of rotors studied by microwave spectroscopists may be classified according to the values of the

TABLE I Types of Molecular Rotors[a]

Spherical top molecules	$I_a = I_b = I_c$	All three moments of inertia are equal, e.g., CH_4
Linear molecules	$I_a = 0,\ I_b = I_c$	Axis a along the internuclear axis, b and c perpendicular to this axis, e.g., OCS, HCl
Symmetric-top molecules		
Prolate top	$I_a < I_b = I_c$	Axis of least moment of inertia, a, along the symmetry axis, e.g., CH_3F
Oblate top	$I_a = I_b < I_c$	Axis of largest moment of inertia, c, along the symmetry axis, e.g., BCl_3
Asymmetric-top molecules	$I_a \neq I_b \neq I_c$	All three moments of inertia different, e.g., SO_2

[a] For a symmetric top, the molecule is designated a prolate or oblate rotor, depending on which inertia axis corresponds to the molecular symmetry axis. Most molecules belong to the asymmetric rotor case.

principal moments of inertia. The various cases are summarized in Table I. It may be noted that molecules with a threefold or higher axis of symmetry are symmetric tops, and this symmetry axis is a principal intertial axis. Also, any two perpendicular axes that are perpendicular to the symmetry axis are principal axes, and the corresponding moments of intertia are equal. As indicated in the table, there are two types of symmetric rotors. For the prolate case, the molecule is elongated like a football, while for the oblate case, the molecule is flattened like a disk. Most molecules are asymmetric tops, and if the molecule has some symmetry, one or more of the principal axes may be selected. If a molecule has a twofold axis of symmetry, then this axis must be a principal axis. If a symmetry plane is present, then two principal axes must lie in this plane and the third must be perpendicular to this plane. It often occurs that two moments of inertia are accidentally close to each other and the slightly asymmetric top approximates one of the symmetric tops. In such cases, it is referred to as a near-prolate or near-oblate asymmetric top.

Expressions for the principal moments of inertia of some simple molecules are collected in Table II. To evaluate the moments of inertia in the general case, an arbitrary but convenient coordinate system may be chosen in the molecule. The center of mass $(\bar{x}, \bar{y}, \bar{z})$ is given by

$$\bar{x} = \frac{\sum m_i x_i'}{M}, \quad \bar{y} = \frac{\sum m_i y_i'}{M}, \quad \bar{z} = \frac{\sum m_i z_i'}{M}, \tag{10}$$

TABLE II Expressions for the Moments of Inertia of Some Simple Molecules

Molecules	Moments of inertia
Diatomic, XY	$I = \left(\dfrac{m_X m_Y}{m_X + m_Y}\right) d_{XY}^2$
Linear, XYZ	$I = \dfrac{1}{M}\left[m_X m_Y d_{YX}^2 + m_Y m_Z d_{YZ}^2 + m_X m_Z (d_{YX} + d_{YZ})^2\right]$
Bent, XY_2[b]	$I_x = 2 m_Y d_{XY}^2 \sin^2\dfrac{\theta}{2}$; $I_y = \dfrac{2 m_X m_Y}{M} d_{XY}^2 \cos^2\dfrac{\theta}{2}$; $I_z = I_x + I_y$
Pyramidal, XY_3[c]	$I_x = I_y = 2 m_Y d_{XY}^2 \sin^2\dfrac{\theta}{2} + \dfrac{m_X m_Y}{M} d_{XY}^2 \times \left(3 - 4\sin^2\dfrac{\theta}{2}\right)$; $I_z = 4 m_Y d_{XY}^2 \sin^2\dfrac{\theta}{2}$

[a] The d_{ij} is the bond distance between atoms i and j; M is the total mass of the appropriate molecule; and m_i is the mass of the ith atom.
[b] The x axis corresponds to the C_2 axis, with the z axis perpendicular to the xy plane and with θ as the Y—X—Y bond angle.
[c] The z axis is the C_3 symmetry axis and θ is the Y—X—Y bond angle. Note that the acute angle β between the X—Y bond and the symmetry axis is related to the bond angle by $\sin(\theta/2) = (\sqrt{3}/2)\sin\beta$.

where x_i', y_i', and z_i' are the coordinates of the ith atoms of mass m_i relative to the arbitrary coordinate system and M is the total mass of the molecule. The center-of-mass coordinates of the atoms are computed from

$$x_i = x_i' - \bar{x}, \qquad y_i = y_i' - \bar{y}, \qquad z_i = z_i' - \bar{z}. \tag{11}$$

The elements of the moment of inertia tensor are evaluated from these coordinates by means of the expressions

$$\begin{aligned} I_{xx} &= \sum m_i\left(y_i^2 + z_i^2\right), & I_{xy} &= -\sum m_i x_i y_i, \\ I_{yy} &= \sum m_i\left(x_i^2 + z_i^2\right), & I_{xz} &= -\sum m_i x_i z_i, \\ I_{zz} &= \sum m_i\left(x_i^2 + y_i^2\right), & I_{yz} &= -\sum m_i y_i z_i. \end{aligned} \tag{12}$$

The $I_{xx}, \ldots$ are the moments of inertia and the $I_{xy}, \ldots$ are called the products of inertia. The inertia matrix $\mathbf{I}$ is symmetric, and diagonalization of this matrix by standard techniques (such as Jacobi's rotation method),

$$\tilde{\mathbf{R}}\mathbf{I}\mathbf{R} = \begin{pmatrix} I_a & 0 & 0 \\ 0 & I_b & 0 \\ 0 & 0 & I_c \end{pmatrix}, \tag{13}$$

provides the principal moments of inertia, and the transformation matrix $\mathbf{R}$ provides the orientation of the principal

axis system relative to the initial arbitary axis system. Equivalently, the secular determinant

$$\begin{vmatrix} I_{xx}-\lambda & I_{xy} & I_{xz} \\ I_{xy} & I_{yy}-\lambda & I_{yz} \\ I_{xz} & I_{yz} & I_{zz}-\lambda \end{vmatrix} = 0 \qquad (14)$$

may be expanded and solved for its roots, which are the principal moments. In general, the above is of third degree in the unknown λ. If, however, one principal axis is known, then by taking one of the initial coordinate axes along this direction, one can make the products of inertia associated with this axis vanish, and the secular equation is simplified.

The masses employed in calculating the moments and products of inertia must correspond to a single isotope for each atom in the molecule. Since the principal moments of inertia are different for different isotopic forms of a molecule, quite different rotational spectra are obtained. In fact, if the molecular mass distribution in a molecule is changed, the rotational spectrum is affected. The spectrum of 2-chloropyridine can, hence, be expected to be quite different from that of 3-chloropyridine. The implications for qualitative analysis of these chemical isomers are obvious. Rotational isomerism also changes the mass distribution, which is illustrated in Fig. 10. The sensitivity of the transition frequency to isotopic composition is shown in Fig. 5 for hydrogen selenide. From a study of the rotational spectrum of different isotopic forms, additional information is obtained that can be used to evaluate detailed molecular structures. Because of the sensitivity of microwave spectroscopy, various isotopic forms can often be studied in natural abundance.

IV. RIGID-ROTOR ENERGY LEVELS AND SPECTRA

To a good approximation, the energy of a molecule may be expressed as the sum of the electronic, vibrational, and rotational energies. In pure rotational spectra, transitions take place between rotational sublevels with no change in the electronic or vibrational state. The gross features of the pure rotational spectrum of a molecule may be ascertained by treating the molecule as a rigid, nonvibrating rotor. Subsequently, the effects of centrifugal distortion and vibration may, in many cases, be included by application of perturbation theory. This is considered further in Sections V and VI. The molecular geometry, via the moments of inertia, determines the pattern of the rigid-rotor spectrum. This pattern is relatively simple for linear and symmetric-top molecules; however, for asymmetric rotors, there is little regularity to the spectrum except in certain cases.

The frequency of electromagnetic radiation absorbed depends on the energy difference between the two states and is given by the Bohr relation

$$\nu = (E_j - E_i)/h, \qquad (15)$$

where h is Planck's constant and E_j and E_i are the upper and lower rotational energy states, respectively. In what

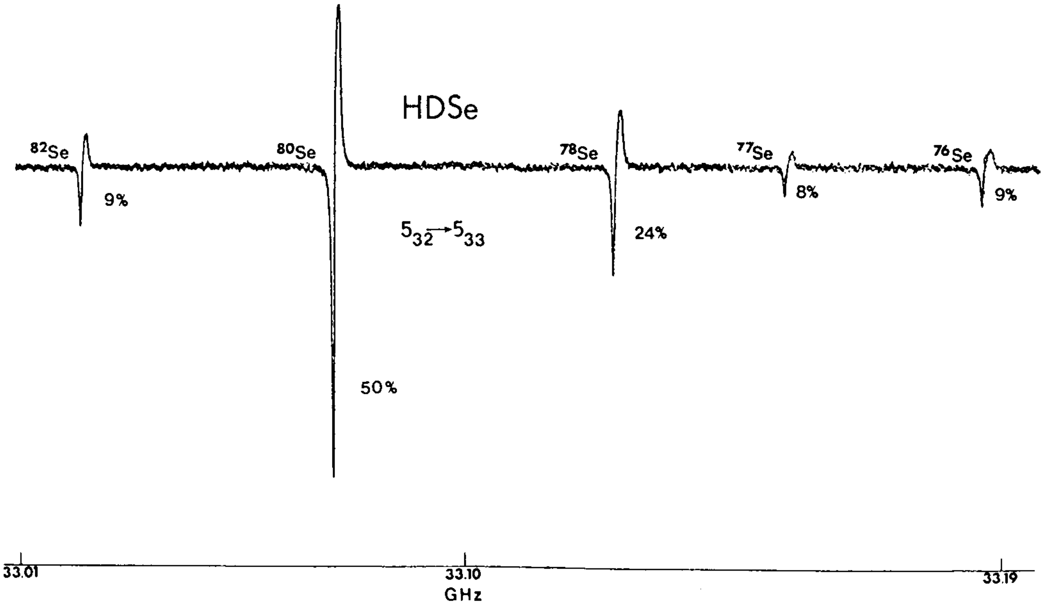


FIGURE 5 A transition of the asymmetric rotor HDSe. Spectral trace was recorded from left to right and extends approximately 170 MHz. The five major isotopic species of Se are clearly evident. The spectrum was taken at the temperature of dry ice; nevertheless, HDSe decomposes in the Stark cell. This is evident from the decrease in line intensity of ^{76}Se compared with that of ^{82}Se.

follows, the factor $1/h$ will be included in the energy expression, so that E will be expressed in frequency units. According to quantum mechanics, gaseous molecules rotate at only certain rates and have discrete values of rotational energy and angular momentum. When a quantum of energy $h\nu$ is absorbed, the molecule is raised to a higher energy level and rotates at the next allowed rate. The type of transition that can occur is governed by the electric dipole moment matrix element or transition moment,

$$\mu_{ij} = \int \psi_i^* \mu \psi_j \, d_\tau = (i|\mu|j), \tag{16}$$

where ψ_i^* and ψ_j are, respectively, the wavefunctions for the lower and upper states of the transition. The i and j denote the set of quantum numbers required to characterize these states. Only those transitions are allowed for which the above matrix element is nonvanishing. This places certain restrictions on the changes possible in the quantum numbers, and these restrictions are called selection rules. The basic problem of rotational spectroscopy is the evaluation of the quantized rotational energy levels and selection rules by quantum mechanical techniques and the subsequent assignment of the observed transitions to particular levels characterized by certain quantum numbers.

The general procedure for evaluating the quantum mechanical energy levels starts with the formulation of the Hamiltonian $\mathcal{H}$, expressed in terms of the angular momentum operators and coordinates, if required. For rigid linear and symmetric-top molecules, the Hamiltonian operator is simple enough that the energy eigenvalue equation may be solved directly,

$$\mathcal{H}\psi = E\psi, \tag{17}$$

to give the eigenfunctions ψ and the energy levels E. For many problems of interest such as an asymmetric top, the above eigenvalue equation cannot be solved directly. In such cases, the wavefunctions may be expressed as linear combinations of a complete set of known functions $\{\phi_n\}$, such as a set of symmetric top functions,

$$\psi = \sum a_n \phi_n. \tag{18}$$

According to quantum mechanics, the energy levels are now found by diagonalization of the Hamiltonian or energy matrix $\mathbf{H} = [\mathcal{H}_{n,n'}]$, where the elements are the matrix elements of $\mathcal{H}$ in the particular basis $\{\phi_n\}$. With the availability of high-speed computers, it is usually a relatively simple matter to diagonalize a symmetric matrix by appropriate numerical methods. Diagonalization of the matrix representation of the operator gives the eigenvalues of the operator

$$\tilde{\mathbf{T}}\mathbf{H}\mathbf{T} = \mathbf{E}, \tag{19}$$

where $\mathbf{E}$ is a diagonal matrix of eigenvalues. The transformation matrix $\mathbf{T}$, which diagonalizes the energy matrix, yields the eigenfunctions. Each column of $\mathbf{T}$ corresponds to the set of expansion coefficients a_n (or eigenvector) for the eigenfunction ψ in the basis $\{\phi_n\}$. These define the set of asymmetric-top functions and are required to calculate the average values of an operator in the asymmetric rotor basis. It may be noted that diagonalization of $\mathbf{H}$ is equivalent to solving the secular determinate

$$|\mathbf{H} - \mathbf{I}E| = \mathbf{0}, \tag{20}$$

which is particularly useful when the order of $\mathbf{H}$ is small. Here $\mathbf{I}$ is a unit matrix.

A. Diatomic and Linear Molecules

The Hamiltonian operator for a rigid diatomic or linear molecule is

$$\mathcal{H} = P^2/2I, \tag{21}$$

where P is the total angular momentum and I the moment of inertia. The operator P_Z corresponds to the projection of P along a space-fixed Z axis. According to quantum mechanics, since the operators $\mathcal{H}$, P^2, and P_Z commute with each other, they have a common set of eigenfunctions, which we denote by $\psi_{J,M} \equiv |J, M)$. Classically, these quantities are constants of motion. The matrix elements of the angular momentum are

$$\begin{aligned} (J, M|P^2|J, M) &= \hbar^2 J(J+1), \\ (J, M|P_Z|J, M) &= \hbar M. \end{aligned} \tag{22}$$

Therefore, the rotational energy is given by

$$E_J = BJ(J+1), \tag{23}$$

where E_J is in frequency units and $B = h/8\pi^2 I$. For linear molecules, the rotational energy levels are characterized by two quantum numbers J and M, which are restricted to certain integral values:

$$J = 0, 1, 2, \ldots, \qquad M = 0, \pm 1, \pm 2, \ldots, \pm J. \tag{24}$$

In the absence of external fields, the rotational energies do not depend on M, as Eq. (23) implies, and all levels are $(2J+1)$-fold degenerate. However, when an external field is applied, this degeneracy is lifted, and the energy depends on the space orientation quantum number M. A similar condition holds for symmetric and asymmetric tops. For rotational absorption of radiation, the selection rule

$$J \to J + 1 \tag{25}$$

applies, and the rotational frequencies of a rigid linear molecule are given by

$$\nu = E_{J+1} - E_J = 2B(J+1). \tag{26}$$

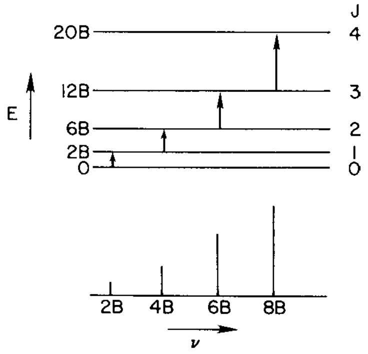


FIGURE 6 The first few rotational energy levels for a linear molecule. The allowed transitions and the resulting spectrum, with approximate intensities, are also shown.

A series of lines at $2B, 4B, 6B, \ldots$ is thus expected for a rigid rotor. The energy levels, allowed transitions, and spectrum of a rigid linear rotor are illustrated in Fig. 6. The molecule OCS, which is commonly used as a standard for various purposes by microwave spectroscopists, has lines that occur at 12,162.97, 24,325.92, 36,488.80, 48,651.40 MHz, ... for the most common isotope. For a light molecule such as CO, $B = 57{,}635.97$ MHz, and the rotational lines are spaced 115,271.94 MHz apart; thus, high-frequency microwave techniques must be employed to measure even the $0 \rightarrow 1$ transition, which is at 115,271.94 MHz. The effect of centrifugal distortion is to produce a small shift to lower frequency in each transition. Illustrative rotational constants are collected in Table III.

TABLE III Spectroscopic Constants and Structures for Some Linear Molecules

Bond length (Å)	B_0 (MHz)	D_0 (kHz)
$\mathrm{H}\overset{1.063}{-}\mathrm{C}\overset{1.155}{\equiv}\mathrm{N}$	44,315.98	87.2
$\mathrm{H}\overset{1.067}{-}\mathrm{C}\overset{1.542}{\equiv}\mathrm{P}$	19,976.01	21.2
$\mathrm{F}\overset{1.262}{-}\mathrm{C}\overset{1.159}{\equiv}\mathrm{N}$	10,554.20	5.3
$\mathrm{F}\overset{1.285}{-}\mathrm{C}\overset{1.541}{\equiv}\mathrm{P}$	5,257.80	1.0
$\mathrm{Cl}\overset{1.629}{-}\mathrm{C}\overset{1.160}{\equiv}\mathrm{N}$	5,970.83	1.7
$\mathrm{Br}\overset{1.789}{-}\mathrm{C}\overset{1.160}{\equiv}\mathrm{N}$	4,120.22	8.8
$\mathrm{N}\overset{1.129}{\equiv}\mathrm{N}\overset{1.189}{=}\mathrm{O}$	12,561.64	5.4
$\mathrm{O}\overset{1.160}{=}\mathrm{C}\overset{1.560}{=}\mathrm{S}$	6,081.49	1.3
$\mathrm{H}\overset{1.053}{-}\mathrm{C}\overset{1.198}{\equiv}\mathrm{C}\overset{1.279}{-}\mathrm{F}$	9,706.19	—
$\mathrm{H}\overset{1.055}{-}\mathrm{C}\overset{1.204}{\equiv}\mathrm{C}\overset{1.637}{-}\mathrm{Cl}$	5,684.24	—
$\mathrm{Na}\overset{1.95}{-}\mathrm{O}\overset{0.96}{-}\mathrm{H}$	12,567.05	28.7
$\mathrm{K}\overset{2.21}{-}\mathrm{O}\overset{0.91}{-}\mathrm{H}$	8,208.68	12.2
$\mathrm{HCN}\overset{2.80}{\cdots}\mathrm{HF}^{a}$	3,591.11	5.2
$\mathrm{OC}\overset{3.07}{\cdots}\mathrm{HF}^{a}$	3,063.90	9.8
$\mathrm{Ar}\overset{3.54}{\cdots}\mathrm{HF}^{a}$	3,065.71	70.9
$\mathrm{Kr}\overset{3.65}{\cdots}\mathrm{HF}^{a}$	2,392.41	31.9
$[\mathrm{H}\overset{1.09}{-}\mathrm{C}\overset{1.11}{\equiv}\mathrm{O}]^{+}$	44,594.42	82.4
$\mathrm{S}\overset{1.4840}{-}\mathrm{O}$	21,523.56	33.9

[a] For these complexes the bridge length r(X—F) is given with X = N, C, Ar, or Kr.

B. Symmetric-Top Molecules

The rotational Hamiltonian for a prolate symmetric top has the form

$$\mathcal{H} = \frac{P^2}{2I_b} + \left(\frac{1}{2I_a} - \frac{1}{2I_b}\right)P_a^2, \tag{27}$$

where $P^2 = P_a^2 + P_b^2 + P_c^2$ is the total angular momentum. In addition to P_Z, a symmetric top has a component of the total angular momentum $P_a(P_Z)$ along the symmetry axis, which is a constant of motion. The quantities $\mathcal{H}$, P^2, P_Z, and P_a commute with each other and hence have a common set of eigenfunctions denoted by $\psi_{JKM} \equiv |J, K, M)$. The matrix elements in the symmetric-top basis,

$$\begin{aligned} (J, K, M \mid P^2 J, K, M) &= \hbar^2 J(J+1), \\ (J, K, M|P_a|J, K, M) &= \hbar K, \\ (J, K, M|P_Z|J, K, M) &= \hbar M, \end{aligned} \tag{28}$$

specify the values of the quantized angular momenta. It follows from Eq. (27) that the energy levels for a prolate rotor are given by

$$E_{J,K} = BJ(J+1) + (A-B)K^2, \tag{29}$$

with the rotational constants defined as $A = h/8\pi^2 I_a$ and $B = h/8\pi^2 I_b$. The energy levels are characterized by the quantum numbers J, K, and M, with

$$\begin{aligned} J &= 0, 1, 2, \ldots, \\ K &= 0, \pm 1, \pm 2, \ldots, \pm J, \\ M &= 0, \pm 1, \pm 2, \ldots, \pm J. \end{aligned} \tag{30}$$

For an oblate top, the unique axis is denoted by c. By replacement of A by C and a by c, the energy expression and angular momentum matrix elements for an oblate top may be obtained. In particular, for the energy,

$$E_{JK} = BJ(J+1) + (C-B)K^2, \tag{31}$$

with $C = h/8\pi^2 I_c$.

As apparent from Eqs. (29) and (31), the energy levels increase with K for a prolate rotor ($A > B$) and decrease with K for an oblate rotor ($C < B$). There are $J+1$ different rotational levels for each J value since the energy does not depend on the sign of K. The rotational levels for $J \leqq 3$ are illustrated in Fig. 7. Furthermore, in the absence of external fields each level is $(2J+1)$-fold degenerate in the space orientation quantum number M. For absorption of radiation, the important selection rules are

$$J \rightarrow J+1, \qquad K \rightarrow K. \tag{32}$$

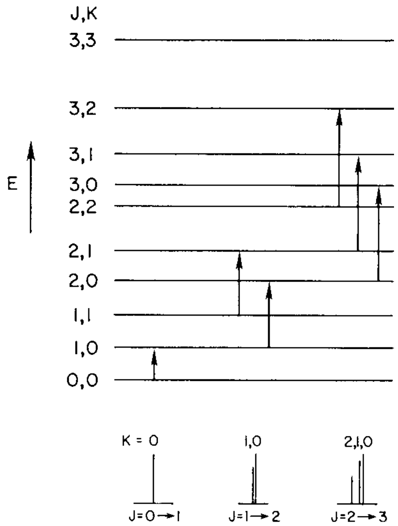


FIGURE 7 Energy levels and allowed $\Delta K = 0$ transitions for a prolate symmetric top. The spectrum is shown at the bottom. Lines of different K for the same J transition are separated by centrifugal distortion.

The rotational line frequencies are, therefore,

$$\nu = 2B(J + 1). \tag{33}$$

The rotational spectrum in this rigid-rotor approximation is like that of a linear molecule shown in Fig. 6. The different $K \to K$ transitions of a given $J \to J + 1$ transition shown in Fig. 7 all have the same frequency. When centrifugal distortion effects are considered, the absorption frequencies are no longer independent of the quantum number K, and this gives rise to a splitting of a given $J \to J + 1$ transition into $J + 1$ closely spaced lines, as shown in Fig. 7. This distinguishes the rotational spectrum of a symmetric top from that of a linear molecule. Table IV summarizes the rotational constants for a few symmetric tops.

Because of the selection rule $\Delta K = 0$, the rotational constant about the symmetry axis, A (C for an oblate top), cannot be evaluated from the rotational spectrum of a symmetric top. This selection rule follows because there is no dipole moment component perpendicular to the symmetry axis, and applies rigorously for a rigid molecule. Recent theoretical developments have shown that actually a very small dipole moment perpendicular to the symmetry axis can be induced through the effects of centrifugal distortion. This lifts the $\Delta K = 0$ selection rule and allows the observation of $\Delta K = \pm 3, \pm 6, \ldots$ transitions with a sensitive spectrometer. This has enabled the structures of pyramidal XY_3 molecules, which have only two structural parameters (see Table II), to be evaluated directly from I_a and I_b without the need for isotopic information. The observation of so-called forbidden transitions often results because of the admixture of rotational states by an interaction. This in turn leads to a nonvanishing transition moment, Eq. (16), which without the mixing of states would vanish.

TABLE IV Spectroscopic Constants and Structures for Some Symmetric-Top Molecules

Molecule	B_0 (MHz)	D_J (kHz)	D_{JK} (kHz)
PH_3	133,480.15	3950	−5180
AsH_3	112,470.59	2925	−3718
CH_3F	25,536.15	59.9	420.3
CH_3CN	9,198.90	3.8	176.9
SiH_3CN	4,973.01	1.5	63
OPF_3	4,811.76	1.0	1.3
$CH_3CN\cdots HF$	1,853.37	0.8	67
$C_6C_6\cdots HCl$	1,237.68	1.2	13.3

Molecule	Bond	Bond length (Å)	Angle	Bond angle (deg)
PH_3	P—H	1.420	HPH	93.3
AsH_3	As—H	1.520	HAsH	92.0
CH_3F	C—H	1.097	HCF	108.4
	C—F	1.384		
CH_3CN	C—H	1.104	HCC	109.4
	C—C	1.458		
	C—N	1.157		
SiH_3CN	Si—H	1.49	CSiH	107.5
	Si—C	1.847		
OPF_3	O—P	1.437	FPF	101.1
	P—F	1.522		
$CH_3CN\cdots HF$	N···F	2.76		
$C_6H_6\cdots HCl$	$B_z\cdots Cl$	3.59		

C. Asymmetric-Top Molecules

The Hamiltonian for an asymmetric top may be expressed as

$$\mathcal{H} = \frac{1}{\hbar^2}\left(AP_a^2 + BP_b^2 + CP_c^2\right), \qquad A > B > C, \tag{34}$$

with $A = h/8\pi^2 I_a$, and so on, in frequency units. $\mathcal{H}$ no longer commutes with P_a, although it still commutes with P^2 and P_Z; hence, J and M are still "good" quantum numbers. The eigenvalue equation for $\mathcal{H}$ cannot be solved directly, but the symmetric-top basis $|J, K, M\rangle$ can be used to construct the energy matrix of $\mathcal{H}$. It is convenient to express $\mathcal{H}$ in terms of an asymmetry parameter. Wang's parameter, particularly appropriate for a near-prolate rotor, is defined by

$$b_p = (C - B)/(2A - B - C), \tag{35}$$

where if $B \cong C$, the parameter is small, $b_p \cong 0$. For a near-oblate top, Wang's parameter is

$$b_0 = (A - B)/(2C - A - B), \tag{36}$$

where if $A \cong B$, $b_0 \cong 0$. An alternate measure of the asymmetry is Ray's parameter

$$\kappa = (2B - A - C)/(A - C). \tag{37}$$

For a prolate symmetric top, $\kappa = -1$, and for an oblate top, $\kappa = +1$. If $B \cong C$, $\kappa \cong -1$, the asymmetric top may be classified as a near-prolate rotor. If $A \cong B$, $\kappa \cong +1$, we have a near-oblate asymmetric rotor. The case $\kappa = 0$ corresponds to the most asymmetric top possible. Most molecules are prolatelike asymmetric tops. Illustrative examples of rotational constants for some asymmetric tops are given in Table V.

In terms of Wang's parameter for a prolate asymmetric rotor, we can reexpress $\mathcal{H}$ as

$$\mathcal{H} = \tfrac{1}{2}(B + C)P^2 + \left[A - \tfrac{1}{2}(B + C)\right]\mathcal{H}(b_p), \tag{38}$$

$$\mathcal{H}(b_p) = P_a^2 + b_p\left(P_c^2 - P_b^2\right). \tag{39}$$

Here and in what follows it is convenient to measure angular momentum in units of $\hbar$, $P_a^2/\hbar^2 \to P_a^2$, and so on. The $\mathcal{H}(b_p)$ is termed a reduced Hamiltonian. In a symmetric rotor representation the nonvanishing matrices of P^2 and $\mathcal{H}(b_p)$ are

$$\begin{aligned}(J, K, M|P^2|JKM) &= J(J + 1)\\ (J, K, M|\mathcal{H}(b_p)|J, K, M) &= K^2\\ (J, K, M|\mathcal{H}(b_p)|J, K \pm 1, M)&\\ = (1/2)b_p\{[J(J + 1) - K(K \pm 1)]&\\ \times [J(J + 1) - (K \pm 1)(K \pm 2)]\}^{1/2}.&\end{aligned} \tag{40}$$

The energy matrix is diagonal in J and has both diagonal and off-diagonal elements in K, the off-diagonal elements coming from the operator $(P_c^2 - P_b^2)$. Since $1/2(B + C)P^2$ contributes only a constant diagonal term, a reduced energy matrix may be defined as

$$E_{K,K'} = (J, K, M|\mathcal{H}(b_p)|J, K', M), \tag{41}$$

with $K' = K$ or $K \pm 1$. Diagonalization of this for each value of J gives the Wang reduced energies $W_{J,\tau}(b_p)$, and the total rotational energy is given by

$$\begin{aligned}E_{J,\tau} &= (1/2)(B + C)J(J + 1)\\ &\quad + [A - (1/2)(B + C)]W_{J,\tau}(b_p).\end{aligned} \tag{42}$$

Expressions for $W(b_p)$ for a few low J levels are given in Table VI. This formulation applies to any asymmetric top; however, if b_p is small, as for a near-prolate rotor, the off-diagonal matrix elements are small, and the matrix is more easily diagonalized. Furthermore, in the case of a near-prolate (or oblate) asymmetric rotor, the $W(b_p)$ differ only slightly from K^2 and may be conveniently expressed

TABLE V Rotational and Centrifugal Distortion Constants[a] and Structures of Some Asymmetric Rotors

Molecule	A	B	C	$\Delta_J(10^3)$	$\Delta_{JK}(10^2)$	$\Delta_K(10^0)$	$\delta_J(10^4)$	$\delta_K(10^2)$
SO_2	60,778.52	10,318.07	8,799.70	6.59	−11.72	2.59	17.01	2.53
S_2O	41,915.44	5,059.10	4,507.16	1.90	−3.19	1.20	3.45	1.22
F_2SO	8,614.80	8,356.95	4,952.94	4.59	−0.25	0.01	15.50	0.22
HCOOH	77,512.23	12,055.10	10,416.15	10.00	−8.63	1.70	19.49	4.26
CH_2═CHF	64,584.69	10,636.88	9,118.03	8.44	−7.59	1.33	17.66	3.57
H_2CO	281,970.52	38,836.04	34,002.20	75.30	129.04	19.41	104.46	102.58
COF_2	11,813.54	11,753.06	5,880.90	6.13	−0.31	0.01	25.78	0.43

Molecule	Bond	Bond length (Å)	Bond	Bond length (Å)	Angle	Bond angle (deg)	Angle	Bond angle (deg)
SO_2	S—O	1.4308			OSO	119.3		
S_2O	S—S	1.884	S—O	1.465	SSO	118.0		
F_2SO	S—O	1.413	S—F	1.585	FSF	92.8	OSF	106.8
HCOOH	C—H	1.097	C═O	1.202	OCO	124.9	H—C═O	124.1
	C—O	1.343	O—H	0.972	COH	106.3		
CH_2═CHF	C—H	1.08	C═C	1.332	HCH	120.6	CCF	121.2
	C—F	1.348			FCH	117.9		
H_2CO	C—H	1.102	C—O	1.204	HCO	121.7		
COF_2	C—F	1.312	C—O	1.174	FCF	108.0		

[a] Constants are multiplied by the given factors of 10; units are MHz.

TABLE VI Wang's Reduced Energies[a]

$J_{K_{-1}K_1}$	$W(b_p)$
$0_{0,0}$	0
$1_{1,0}$	$1 - b_p$
$1_{1,1}$	$1 + b_p$
$1_{0,1}$	0
$2_{2,0}$	$2 + 2\sqrt{1 + 3b_p^2}$
$2_{2,1}$	4
$2_{1,1}$	$1 - 3b_p$
$2_{1,2}$	$1 + 3b_p$
$2_{0,2}$	$2 - 2\sqrt{1 + 3b_p^2}$

[a] Total asymmetric-top energy is given by Eq. (42) for any asymmetry b_p.

as an expansion in terms of Wang's asymmetry parameter. The energy matrix can actually be factored into four smaller submatrices by taking as basis functions a set of functions that are linear combinations of the original functions: $(\psi_{JKM} \pm \psi_{J-K,M})/\sqrt{2}$ and ψ_{JOM}. These functions may be classified according to the symmetry of the rotational Hamiltonian, the four-group, which has the symmetry species A, B_a, B_b, and B_c. The form of the submatrices is tridiagonal:

$$E^+(\kappa) = \begin{bmatrix} E_{00} & \sqrt{2}E_{02} & 0 & 0 \\ \sqrt{2}E_{02} & E_{22} & E_{24} & 0 \\ 0 & E_{24} & E_{44} & \cdot \\ 0 & 0 & \cdot & \cdot \end{bmatrix},$$

$$E^-(\kappa) = \begin{bmatrix} E_{22} & E_{24} & 0 & 0 \\ E_{24} & E_{44} & E_{46} & 0 \\ 0 & E_{46} & E_{66} & \cdot \\ 0 & 0 & \cdot & \cdot \end{bmatrix},$$

$$0^+(\kappa) = \begin{bmatrix} E_{11} + E_{-11} & E_{13} & 0 & 0 \\ E_{13} & E_{33} & E_{35} & 0 \\ 0 & E_{35} & E_{55} & \cdot \\ 0 & 0 & \cdot & \cdot \end{bmatrix}, \quad (43)$$

$$0^-(\kappa) = \begin{bmatrix} E_{11} - E_{-11} & E_{13} & 0 & 0 \\ E_{13} & E_{33} & E_{35} & 0 \\ 0 & E_{35} & E_{55} & \cdot \\ 0 & 0 & \cdot & \cdot \end{bmatrix}.$$

with the matrix elements calculated from Eq. (41), taking cognizance of Eqs. (40) and (44). Since the matrices are Hermitian, $(K|\mathcal{H}|K') = (K'|\mathcal{H}|K)^*$, we have the relations

$$E_{K,K} = E_{-K,-K}, \qquad E_{K,K+2} = E_{K+2,K} = E_{-K,-K-2} = E_{-K-2,-K}. \quad (44)$$

The size of the above submatrices depends on the maximum value of K, which in turn depends on the value of J being considered. For an oblatelike asymmetric rotor ($\kappa > 0$), A and C as well as a and c are interchanged in the above relations:

$$E_{J,\tau} = (1/2)(A + B)J(J + 1) + [C - (1/2)(A + B)]W_{J,\tau}(b_o). \quad (45)$$

A parallel formulation using κ can also be made and has often been used in the literature. Here the reduced energies are denoted by $E_{J,\tau}(\kappa)$, and the total energy is given by

$$E_{J,\tau} = (1/2)(A + C)J(J + 1) + (1/2)(A - C)E_{J,\tau}(\kappa). \quad (46)$$

The reduced energies are again found by diagonalization of the matrices of Eq. (43), with the specific elements now a function of κ.

The effect of the asymmetry is to remove the twofold degeneracy of the K levels ($K \neq 0$) for a symmetric top. This K-splitting of the levels increases as the asymmetry increases. There are thus $2J + 1$ different rotational levels or values of $W(b_p)$ for each value of J. To distinguish the sublevels for a given J, the notation J_τ is employed, where τ takes on $2J + 1$ integer values ranging from $-J$ to $+J$. These are assigned to the energy levels such that $\tau = -J$ corresponds to the lowest energy, $\tau = -J + 1$ to the next lowest energy level, and finally $\tau = J$ to the highest level. The energy levels for an asymmetric rotor are thus characterized by the quantum numbers J and M and by the pseudo-quantum number τ. An alternate notation is more frequently employed where two subscripts are added to J, that is, J_{K_{-1},K_1}. A particular asymmetric-top level converges to a particular prolate symmetric-top level specified by $|K|$ as κ approaches -1 and to a particular oblate symmetric-top level as κ approaches $+1$. The subscript K_{-1} is therefore the $|K|$ value of the prolate rotor limit, $\kappa = -1$, and K_1 is the $|K|$ value of the oblate rotor limit, $\kappa = 1$. The relation between the two notations is

$$\tau = K_{-1} - K_1. \quad (47)$$

The possible energy levels for $J = 1$ may hence be denoted as 1_{-1}, 1_0, 1_1, or more commonly as $1_{0,1}$, $1_{1,1}$, $1_{1,0}$. The first few rotational energy levels of an asymmetric top are illustrated in Fig. 8.

For dipole absorption of radiation, the selection rules for J are as follows, along with their general designations:

$$\begin{aligned} &J \to J + 1, && \text{R-branch transition,} \\ &J \to J, && \text{Q-branch transition,} \\ &J \to J - 1, && \text{P-branch transition.} \end{aligned} \quad (48)$$

The transitions $\Delta J = 0$, $\Delta J = -1$ can give rise to absorption lines for an asymmetric rotor, although not for a rigid

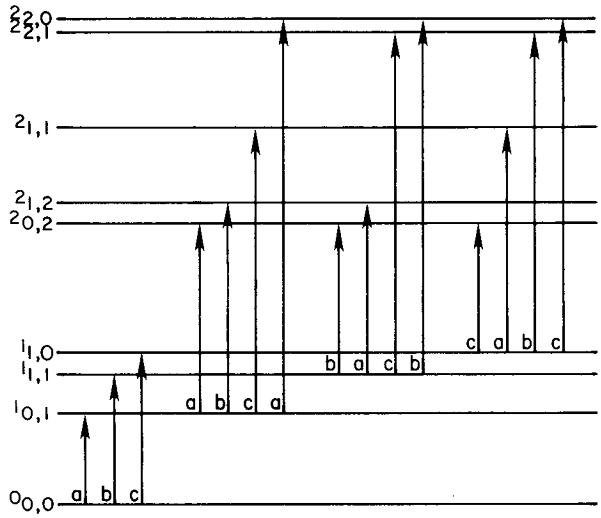


FIGURE 8 Schematic of the lower energy levels of an asymmetric top. The allowed *a*-, *b*-, and *c*-type, $\Delta J = +1$ transitions are given.

linear or symmetric rotor. The $\Delta J = -1$ transitions arise because not all levels are always ordered by their J values. The selection rules for K_{-1} and K_1 depend on the possible components of the molecular dipole along the principal inertial axes. In general, an asymmetric top can have a dipole moment component along all three principal axes. The dipole component along the axis associated with the least, the intermediate, and the greatest moment of inertia are denoted μ_a, μ_b, and μ_c, respectively. Transitions that arise from the μ_a component are designated as a-type transitions. Similarly, those due to μ_b or μ_c are designated respectively as b-type or c-type transitions. If, for example, $\mu_a = 0$, then no a-type transitions are allowed. The selection rules for the subscripts K_{-1} and K_1 can be expressed in terms of the allowed changes of K_{-1} and K_1 or in terms of the evenness or oddness of the subscripts. If K_{-1} is an even (odd) integer, it is designated e (o); similarly for K_1. The resulting selection rules are summarized in Table VII.

TABLE VII Selection Rules for K_{-1} and K_1

Direction of dipole component	Permitted transitions				
	ΔK_{-1}	ΔK_1	$K_{-1}K_1$	$\leftrightarrow$	$K'_{-1}K'_1$
μ_a	$0, \pm 2, \ldots$	$\pm 1, \pm 3, \ldots$	ee	$\leftrightarrow$	eo
			oe	$\leftrightarrow$	oo
μ_b	$\pm 1, \pm 3, \ldots$	$\pm 1, \pm 3, \ldots$	ee	$\leftrightarrow$	oo
			oe	$\leftrightarrow$	eo
μ_c	$\pm 1, \pm 3, \ldots$	$0, \pm 2, \ldots$	ee	$\leftrightarrow$	oe
			eo	$\leftrightarrow$	oo

A transition such as $1_{0,1} \rightarrow 2_{0,2}$, is hence an R-branch, a-type transition, while the transition $2_{1,1} \rightarrow 2_{2,0}$ is a Q-branch, b-type transition. Levels are also often referred to in terms of their symmetry under the four-group. For example, when $K_{-1}K_1$ is ee, eo, oe, or oo, the levels, respectively, belong to the symmetry species A, B_a, B_c, or B_b of the four-group. The allowed a-type transitions in this notation are $A \leftrightarrow B_a$ and $B_c \leftrightarrow B_b$. The allowed $\Delta J = +1$ transitions are indicated in Fig. 8 for $J \leqq 2$.

As an approximate rule, the most intense lines are those for which $\Delta K_{-1} = 0, \pm 1$ for $\kappa \cong -1$ and $\Delta K_1 = 0, \pm 1$ for $\kappa \cong +1$. Unlike linear and symmetric-top molecules, transitions between high-J levels can fall at low frequencies for an asymmetric top. Because of the greater number of possible transitions and the lack, in many cases, of any regular spectral pattern, the rotational spectrum of an asymmetric rotor is more complicated and more difficult to assign. Various effects, such as the Stark effect, have been found to be useful aids in the assignment of such spectra. A rotational spectrum of an asymmetric rotor is given in Fig. 9. The spectrum is a moderately rich spectrum, and the lack of any obvious pattern is apparent. In some situations, spectra with a rather regular pattern are obtained. For a-type transitions ($\Delta J = 1$, $\Delta K_{-1} = 0$) of a very near prolate rotor ($\kappa \cong -1$), a spectral pattern that closely resembles that of a symmetric top is obtained. For small asymmetry, $W(b_p) \cong K^2$, and Eq. (42) gives for the absorption frequencies

$$\nu \cong (B + C)(J + 1). \qquad (49)$$

What will be observed is a group, or cluster, of lines centered approximately at the above frequency. Each group of lines is separated from the next group by approximately $(B + C)$. A similar situation occurs for the c-type transitions of a very near oblate rotor ($\kappa \cong 1$); the separation is now approximately $(A + B)$. Such a spectrum is shown in Fig. 10. Note the regular separation between the absorption bands in this low-resolution spectrum. This figure illustrates the sensitivity of microwave spectroscopy to rotational isomerism. Two sets of band spectra are observed that arise from the two different isomers of crotonic acid. It should be noted that the actual appearance of the spectrum depends on the scan rate, Stark voltage, and sample pressure. Under slower scan rates, individual lines may be observed within each absorption band.

D. Spectral Assignment Aids

Identification of an observed rotational transition with the corresponding rotational quantum numbers of the two levels involved in the transition is termed spectral assignment. The assignment of the ground vibrational state of linear and symmetric rotors is relatively straightforward because

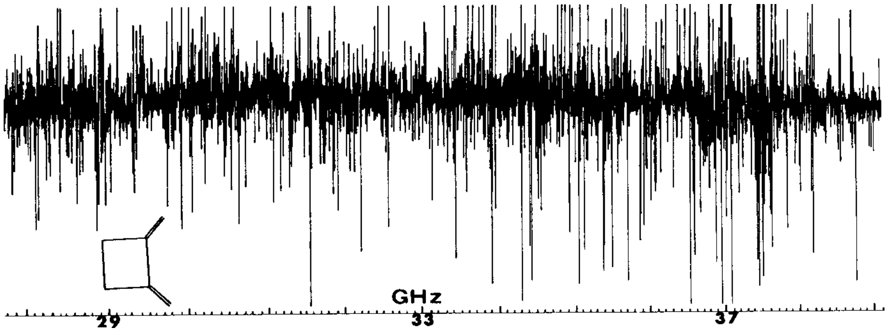


FIGURE 9 Low-resolution survey scan of the R-band spectrum for 1,2-dimethylenecyclobutane. Stark field ~3000 V/cm, pressure ~80 μm, and temperature approximately that of dry ice. Stark lobes are up and zero-field lines are down.

the spectral patterns, as illustrated by Figs. 6 and 7, are simple and easily recognizable. In excited vibrational states, additional effects can arise (see Section VI), which complicate the assignment. On the other hand, the spectrum of an asymmetric rotor, even in the ground vibrational state, can be quite complex, as illustrated in Fig. 9.

The analysis of any spectrum starts with the evaluation of the moments of inertia from a trial structure obtained, for example, from similar molecules. Next, estimates are made of the dipole components μ_a, μ_b, and μ_c from bond dipole moments. These, along with the selection rules, indicate the types of transitions that will dominate the spectrum. This information and the moments of inertia enable the rotational spectrum to be properly predicted. The predicted and observed spectra can then be compared. Various tentative assignments of the measured lines can be made until a correct assignment is found which accounts for the details of the measured spectrum. In this way, the rotational constants are evaluated which characterize the molecular structure.

There are a number of useful aids which can help in establishing and confirming assignments. The Stark effect, discussed in Section IX, can give Stark effect patterns which, when resolvable, may be used to distinguish a Q-branch from either a P- or R-branch transition and identify the lower J value of the transition. If the molecule exhibits nuclear quadrupole splitting (see Section X), certain transitions can exhibit characteristic splitting patterns which can be helpful in assigning the spectrum. Spectral patterns, particularly for slightly asymmetric rotors, can provide useful information for assignments. For example, the a-type, R-branch transitions of a near-prolate asymmetric rotor, $J_{K_{-1},K_1} \rightarrow (J+1)_{K_{-1},K_1+1}$, give rise to a closely spaced cluster of lines for a given $J \rightarrow J+1$. Except for $K_{-1}=0$, there are two transitions for each K_{-1} $(2J+1$ lines). Within a given cluster the intensity of $K_{-1} \rightarrow K_{-1}$ lines decreases with increasing K_{-1} because of the Boltzmann factor in the intensity expression. Also, the lower K_{-1} doublet transitions are more widely split by the inertial asymmetry than the higher K_{-1} lines.

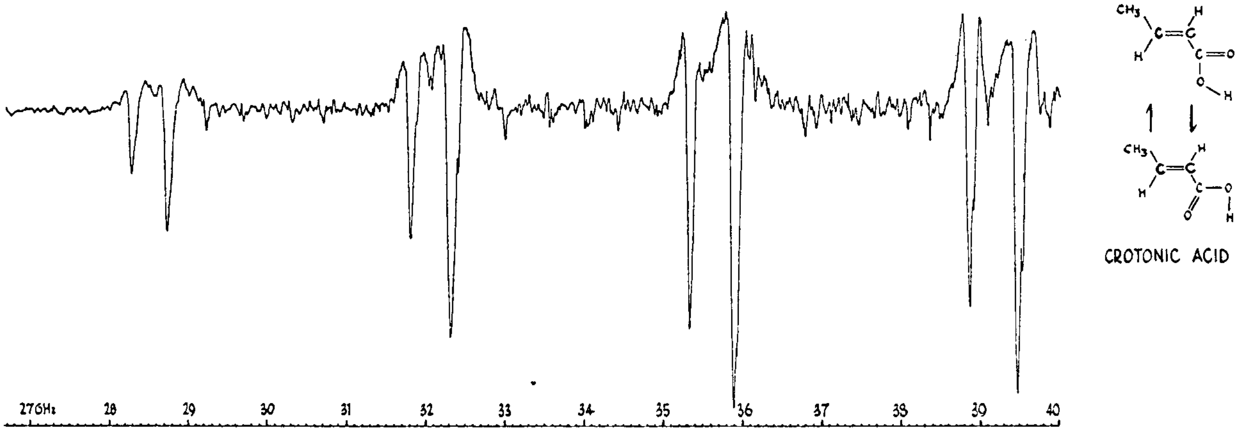


FIGURE 10 Low-resolution microwave spectrum of crotonic acid. Scan rate 10 MHz/sec, pressure about 15 mTorr, 1000-V Stark modulation. Spectra of two rotational isomers (s-*trans* and s-*cis*) are evident. Strongest bands are due to s-*trans*. [Courtesy of Hewlett-Packard Co.]

Often only the two $K_{-1}=1$ lines are widely split, while the remaining lines bunch together between the $K_{-1}=1$ lines. Under moderate resolution, it is often possible to identify this pattern and assign the $K_{-1}=1$ lines. Under low resolution, bandlike spectra are obtained as illustrated in Fig. 10. The separation of the bands yields $(B+C)$, which provides a good start to the assignment process.

A more recent aid to the assignment of rotational spectra is the application of double resonance discussed in the next section.

1. Double Resonance

In the double-resonance technique, two radiation fields, a pump (ν_p) and a signal (ν_s), are introduced simultaneously into the waveguide absorption cell. The pump radiation is high powered, while the signal is low powered. When both the pump and signal are in the microwave region, the method is known as microwave–microwave double resonance (MMDR). The process can be explained qualitatively by reference to the three-level energy system of Fig. 11. The effect of the high-powered pump ν_p $(1\rightarrow 2)$ is to reduce the population of state 1 and hence decrease the intensity of the absorption signal ν_s $(1\rightarrow 3)$. On the other hand, if the observed transition ν_s is the $2\rightarrow 3$ transition (not depicted), then ν_p $(1\rightarrow 2)$, which raises the population of state 2, would enhance the observed absorption signal. Hence, the pump radiation clearly affects the microwave absorption signal by virtue of the fact the transitions share a common energy level. Knowledge of these energy level connections provides a basis for spectral assignments.

Spectrometers have been built around this phenomenon to simplify complicated spectra. For example, instead of Stark modulation, the pump source may be modulated, and this modulatates the signal absorption by means of the double-resonance effect. A narrowband amplifier tuned to the modulation frequency is employed along with the usual phase-sensitive detector. A simplified version of such a spectrometer is shown in Fig. 12. Clearly, as the signal frequency is swept, only transitions with a common energy level with the fixed pump will be modulated and observed, thus greatly simplifying the spectrum. For proper operation of the spectrometer, a filter is used to prevent the modulated pump radiation from reaching the signal detector.

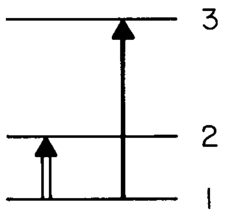


FIGURE 11 An example of a three-level system showing dipole-allowed transitions.

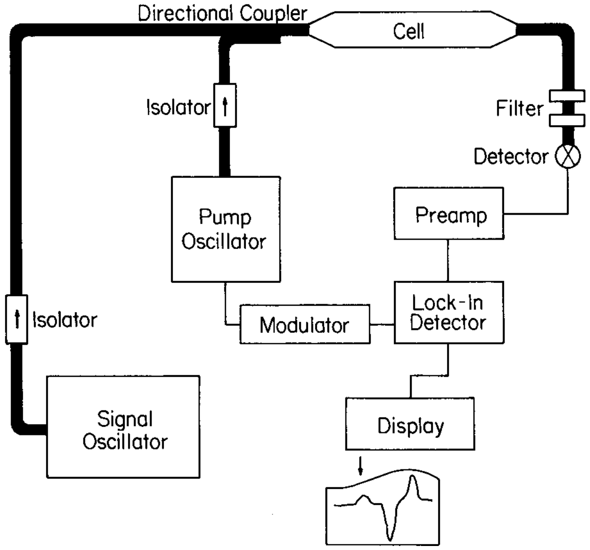


FIGURE 12 A simplified MMDR spectrometer.

When the energy levels for the pump transition are closely spaced, the pump frequency can be in the radiofrequency region, and radiofrequency–microwave double-resonance (RFMDR) experiments can be employed to simplify the spectrum. The K-doublets of slightly asymmetric rotors are often split by a few megahertz and lie in the RF region. Typically, a conventional Stark-modulated spectrometer is used, and the strong radiofrequency source is square-wave modulated and applied across the Stark cell instead of the usual Stark modulation. Without special impedance matching, the upper limit for the RF radiation is about 100 MHz. Again, only signal transitions will be observed which share a common energy level with the pump and are therefore modulated by the pump radiation.

Other types of experiments based on double-resonance techniques provide alternate useful information as outlined in Table VIII. Moreover, double-resonance techniques have been used with optical and infrared radiation to observe microwave transitions from excited electronic states and excited vibrational states, respectively.

TABLE VIII Information from Microwave–Microwave Double Resonance

Energy level structure
Spectral assignments
Relaxation phenomena, rotational relaxation
Rotational energy transfer processes
Collision—induced transitions
Collisional selection rules
Enhancement of weak or forbidden transitions

a. Rotational energy transfer. As indicated above, rotational energy transfer can be studied by use of double-resonance techniques. For example, by using infrared pump radiation with probe radiation in the millimeter- or submillimeter-wave region, both rotational and vibrational state-changing collisions can be studied. Here, a CO_2 laser coincidence with an infrared molecular transition provides a nonthermal population distribution in a particular rotational state of the excited vibrational state. The infrared radiation can be introduced into the measurement cell along with the microwave probe radiation via a dichroic window. The probe radiation probes a single rotational transition within the excited vibrational state. The change in intensity of the rotational line, or time response, is recorded after the pumping laser is pulsed. The probe hence monitors the population change as the system is perturbed and returns to equilibrium.

In the case of studies on the symmetric top CH_3F, for example, the CO_2 laser line 9P(20) pumps the CH_3F energy level population from $J = 12$, $K = 2$ of the ground vibrational state to the $v_3 = 1$, $J = 12$, $K = 2$ state. The excess population in the excited vibrational state produced by the pump is quickly redistributed by several CH_3F—CH_3F collisional processes to reestablish rotational thermal equilibrium within $v_3 = 1$. Also, several vibrational collision processes reestablish equilibrium among the vibrational levels. The measured time-varying responses for the various rotational transitions provide data to characterize the complicated internal energy transfer. Both J-changing rotational collisions, where $v_3 = 1$, $\Delta J = n$, $\Delta K = 0$ (n integer), and K-changing rotational collisions, where $v_3 = 1$, $\Delta J = m$, $\Delta K = 3n$ (m, n integers) have different behavior but can be accurately modeled with four parameters.

Many more studies on rotational energy transfer can be expected in the future. Pressure broadening effects observed in rotational lines due to self-broadening or foreign gas broadening also provide information on energy transfer processes during molecular collisions; however, the information is much less than the method described here. Various pressure broadening studies have been carried out over the years. Pressure broadening is discussed further in Section XI.

2. *Ab Initio* Quantum Calculations

High-quality *ab initio* calculations have proven to be a helpful aid in the analysis of rotational spectra, particularly for predictive purposes and where there exists ambiguities. In general, *ab initio* calculations can provide useful information for structural determinations. These include (i) estimates of the most stable conformers and the lowest energy form, (ii) insight into the difference expected in structural parameters between conformers, (iii) a guide to which parameters are expected to differ significantly between conformers, or to unique structural effects, (iv) a basis to fix particular bond distances or angles or their differences in the structural analysis, (v) initial structural parameters for prediction of spectra of new and unusual molecules, (vi) calculation of quadratic force constants for the estimation of quartic distortion constants or the evaluation of average structures, (vii) estimation of the dipole components μ_a, μ_b, and μ_c from which transition intensities may be computed, and (viii) calculation of the cubic force field to evaluate the rotation–vibration interaction constants α_i and subsequently equilibrium rotational constants from the effective rotational constants.

The computation of semiexperimental equilibrium rotational constants from calculated cubic force constants allows the important equilibrium structure to be obtained. Results of this approach have been quite satisfactory and this can be expected to be a very useful approach where limited vibrational data are available.

In the case of weakly bound complexes, *ab initio* calculations can be very useful to limit the number of possible configurations that need to be considered in the structural determination. Furthermore, the *ab initio* results can help to remove other ambiguities which arise and to identify possible large-amplitude motions.

In addition, various microwave-derived internal rotation barriers and fine structure and hyperfine structure coupling constants all provide a wealth of data to test the accuracy of quantum mechanical calculations.

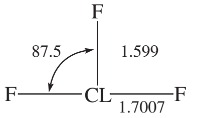
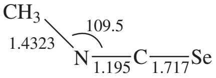
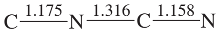
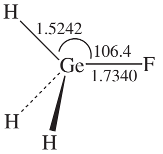
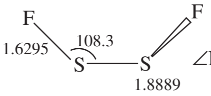
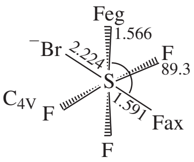
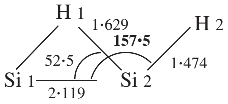
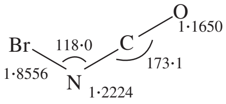
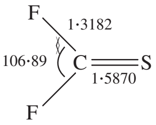
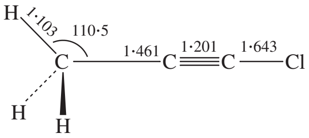
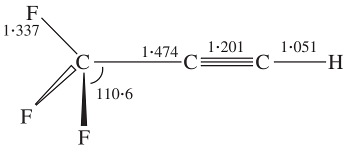
E. Structural Information

Structures of many hundreds of molecules—diatomic, linear, symmetric top, and asymmetric top—have been evaluated by microwave spectroscopy. Illustrative structures are given in Tables III–V and XV (see also Section VIII). Structures of some small molecules are given in Table IX.

Over the last 10 years, a considerable amount of work has been devoted to the study of ions, radicals, molecular complexes, and other transient molecular species due primarily to improved experimental techniques, which are discussed in Section XI. Results of some of these studies will be addressed in this section. Structures of a few unstable molecules along with methods for their preparation are given in Table X. Methods of evaluating molecular structures and the complications arising from effects of molecular vibration are discussed in Section VIII.

By means of the very precise structural parameters that can be obtained from microwave spectroscopy investigations, trends in bond distances or angles among a series of structurally similar molecules can be readily studied. At present the most reliable parameters for many large molecules are those derived by the substitution method

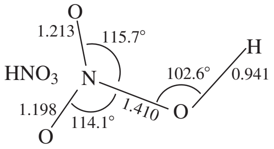
TABLE IX Illustrative Structures of Some Small Molecules Derived by Microwave Spectroscopy[a]

Structure	Reference
$[r_0]$ 	Haubrich, S. T., Roehrig, M. A., and Kukolich, S. G. (1990). *J. Chem. Phys.* **93,** 121
$[r_0]$  C—H = 1.100	Koput, J., Stroh, F., and Winnewisser, M. (1990). *J. Mol. Spectrosc.* **140,** 31
$[r_s]$ 	Gerry, M. C. L., Stroh, F., and Winnewisser, M. (1990). *J. Mol. Spectrosc.* **140,** 147
$[r_e]$ 	Le Guennec, M., Chen, W., Wlodarczak, G., Demaison, J., Eujen, R., and Burger, H. (1991). *J. Mol. Spectrosc.* **150,** 493
$[r_e]$  ∠FS—SF = 87.5	Davis, R. W., and Firth, S. (1991). *J. Mol. Spectrosc.* **145,** 225
$[r_0]$ 	Brier, P. N. (1991). *J. Mol. Struct.* **263,** 133
$[r_0]$ 	Cordonnier, M., Bogey, M., Demuynck, C., and Destombes, J. L. (1992). *J. Chem. Phys.* **97,** 7984
$[\langle r \rangle]$ 	Hensel, K. D., Lam, M. E., Gerry, M. C. L., and Willner, H. (1992). *J. Mol. Spectrosc.* **151,** 184
$[\langle r \rangle]$ 	Xu, Y., Gerry, M. C. L., Joo, D. L., and Clouthier, D. J. (1992). *J. Chem. Phys.* **97,** 3931
$[\langle r \rangle]$ 	Cox, A. P., Ellis, M. C., and Perrett, T. (1992). *J. Chem. Soc. Faraday Trans.* **88,** 2611
$[\langle r \rangle]$ 	Cox, A. P., Ellis, M. C., Legon, A. C., and Wallwork, A. (1993). *J. Chem. Soc. Faraday Trans.* **89,** 2937

continues

TABLE IX (*Continued*)

[⟨r⟩]

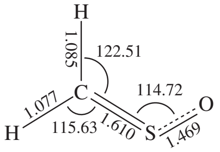


Cox, A. P., Ellis, M. C., Attfield, C. J., and Ferris, A. C. (1994). *J. Mol. Struct.* **320,** 91

[r_e]

Esposti, C. D., Tamassia, F., Cazzoli, G., and Kisiel, Z. (1995). *J. Mol. Spectrosc.* **170,** 582

[r_s]

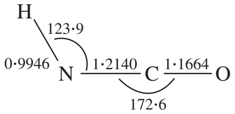


Klesing, A., and Sutter, D. H. (1995). *J. Mol. Struct.* **352,** 357

[r_e]

Nakata, M., and Kuchitsu, K. (1995). *J. Mol. Struct.* **352,** 219

[r_s]



Niedenhoff, M., Yamsda, K. M. T., Winnewisser, G., and Ross, S. C. (1995). *J. Mol. Struct.* **352,** 423

[a] Type of structure (see Section VIII) is given in brackets. Bond lengths are in angstrom units, bond angles ABC in degrees. See also references cited.

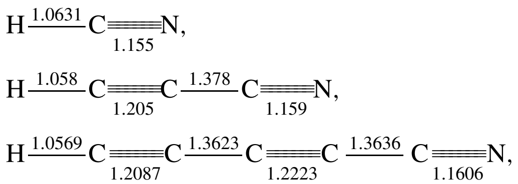
and the scaled method, and critical evaluation of predictive models should make use of these parameters if possible.

Numerous small structural effects have also been revealed by microwave structure studies. In this section a few of these results are pointed out to provide the reader with an indication of what can be found from a study of the rotational spectrum. Also, some of the areas of study are mentioned briefly to indicate the wide range of problems being attacked by microwave spectroscopy. Many detailed structures and molecular information other than those mentioned here have been obtained, and the reader should consult the Bibliography and references cited therein for further examples.

1. Selected Structural Information

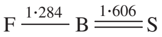
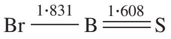
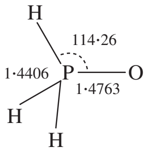
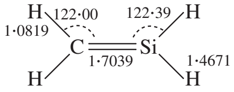
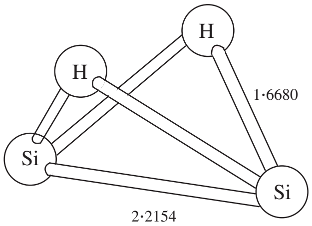
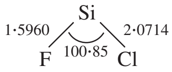
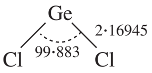
A careful study of the molecular structure by microwave methods provides numerous interesting structural features. For example, it is apparent from Table III that the CF bond in FCN is shorter than in FCP. Many observations of this nature have been made and often correlated with simple bonding theory. A different effect may be illustrated by the —C≡N chain in methylene cyanide, $CH_2(CN)_2$, and sulfur dicyanide, $S(CN)_2$, which is not linear, but rather is bent slightly by about 3° and 5°, respectively. Such effects are not confined to the presence of a cyano group. Similar results are obtained for the acetylene moiety in $H_2PC{\equiv}CH$ depicted in Fig. 13a. Here, the C≡CH group is bent away from the phosphine hydrogens by 7°. A small nonlinearity in the NCO, NCS, and N_3 chains is found in HNCO, HNCS, HN_3, and ClN_3. Many other examples of slightly bent configurations have been found. Some are not easily explained by simple bonding theory.

Recent microwave measurements in the laboratory have aided the detection of many molecules in interstellar space. We mention here the linear cyanide chains,



which have been studied in the laboratory and also detected in interstellar space. Such observations in outer space provide the information from which theories on the mechanisms of formation of particular molecules can be formulated and hence a better understanding gained of the development of the universe. The structures of the above molecules were obtained by studying various ^{13}C isotopic species as well as ^{15}N and D species. Ideally, each nonequivalent atom needs to be isotopically substituted

TABLE X Structures and Preparation Methods of Some Unstable Molecules[a]

Structure	Reference
Linear[b] 	Cooper, T. A., Firth, S., and Kroto, H. W. (1991). *J. Chem. Soc. Faraday Trans.* **87,** 1499
Linear[b] 	Cooper, T. A., Firth, S., and Kroto, H. W. (1991). *J. Chem. Soc. Faraday Trans.* **87,** 1
Symmetric top[c] 	Ahmad, I. K., Ozeki, H., and Saito, S. (1991). *J. Chem. Phys.* **110,** 912
Planar[d] 	Bailleux, S., *et al.* (1997). *J. Chem. Phys.* **106,** 10016
Debridged[e] (butterfly) ∠HSi—SiH = 104.22 	Bogey, M., Bolvin, H., Cordonnier, M., Demuynck, C., Destombes, J. L., and Csaszar, A. G. (1994). *J. Chem. Phys.* **100,** 8614
Planar[f] 	Fujitake, M., and Hirota, E. (1997). *J. Mol. Struct.* **413,** 21
Planar[g] 	Tsuchiya, M. J., Honjou, H., Tanaka, K., and Tanaka, T. (1995). *J. Mol. Struct.* **352,** 407

[a] Bond lengths in angstrom units, bond angles in degrees.
[b] Prepared by passing FSSF or BrSSBr vapor over crystalline boron chips at ca. 1100°C.
Reaction product introduced directly into the microwave cell and rapidly pumped through the cell (fast pumping) [r_s].
[c] Prepared *in situ* by a dc glow discharge of a mixture of PH_3, CO_2, and H_2 [r_0].
[d] Prepared by a pyrolysis reaction, equilibrium structure [r_e].
[e] Prepared by dc glow discharge of SiH_4 in Ar [$\langle r \rangle$].
[f] Prepared by fast pumping using pyrolysis on a mixture of compounds [$\langle r \rangle$].
[g] Prepared by fast pumping and a dc glow discharge of $GeCl_4$ [r_e].

(see Section VIII). The isotopic species data provide the Cartesian coordinates of the atoms in the molecule's principal axis system. For HC_5N, the spectra of eight isotopic species were studied in natural abundance. The above series also nicely illustrates delocalization effects in such conjugated systems. In going from HC_3N to HC_5N, delocalization lengthens the triple bonds and shortens the single bonds.

Other members of these linear carbon chains have also been measured. The unstable linear carbon chain molecule HC_9N,

$$H{-}C{\equiv}C{-}C{\equiv}C{-}C{\equiv}C{-}C{\equiv}C{-}C{\equiv}N,$$

was first detected in interstellar space. It was thought that this molecule would be very difficult to produce and measure in the laboratory. However, it has now been observed in the laboratory by discharging a mixture of vinylcyanide and acetylene diluted in argon followed by rapid expansion into a cavity spectrometer which stabilizes the molecule.

The methylcyanopolyynes, $CH_3(C{\equiv}C)_nCN$, like the cyanopolyynes, $H(C{\equiv}C)_nCN$, are of interest both for possible detection in interstellar space and for testing

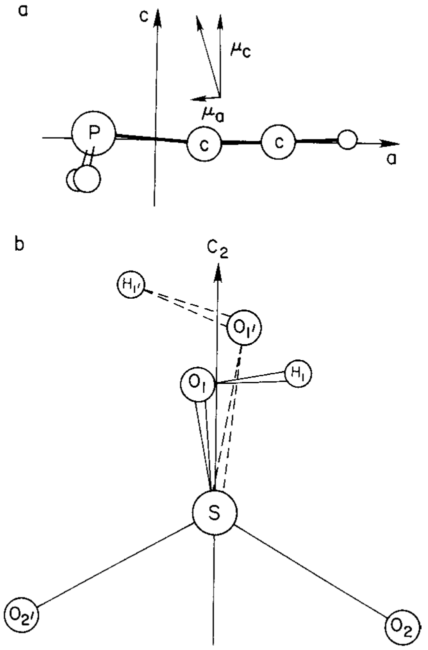


FIGURE 13 (a) Structure, orientation of principal axis, and dipole moment components in $H_2PC{\equiv}CH$. [From Cohen, E. A., McRae, G. A., Goldwhite, H., Stefano, S. D., and Beaudet, R. A. (1987). *Inorg. Chem.* **26,** 4000.] (b) Conformation of sulfuric acid showing the C_2 axis. [From Kuczkowski, R. L., Suenram, R. D., and Lovas, F. J. (1981). *J. Am. Chem. Soc.* **103,** 2561.]

ab initio quantum calculations. The structural formulas through $n = 5$ are given in Fig. 14. The $n = 0, 1$ molecules have been detected in a cold molecular cloud. The molecules were studied by pulsed microwave Fourier transform (MWFT) spectroscopy. A low-current dc discharge of methylcyanoacetylene in Ne was used to produce methylcyanopolyynes to $n = 4$. The rotational constants of Fig. 14 provide the basis for predicting radio transitions of interest for astronomical searches. Similarly, a recent study has provided accurate rotational

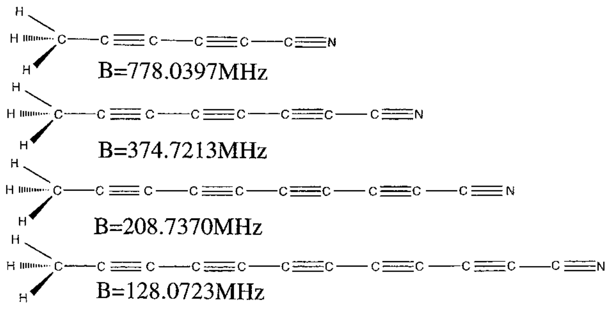


FIGURE 14 Methylcyanopolyyne structural formulas and microwave-determined rotational constants. For details of the analysis see Chen, W., Grabow, J.-U., Travers, M. J., Munrow, M. R., Novick, S. E., McCarthy, M. C., and Thaddeus, P. (1998). *J. Mol. Spectrosc.* **192,** 1.

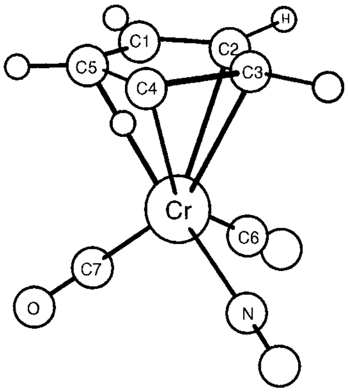
constants for two isocyanopolyynes, $H{-}(C{\equiv}C)_2NC$ and $H{-}(C{\equiv}C)_3NC$.

Another illustration is provided by the common substance sulfuric acid, $(HO)_2SO_2$. The absorption spectrum of discrete sulfuric acid molecules has been recently studied by microwave spectroscopy. A detailed molecular structure determination indicates the free acid does not have the expected C_{2v} symmetry, but rather the symmetry is only C_2. The stable conformation is depicted in Fig. 13b. The interesting features found, which show the power of the method, are that the OH groups are rotated ca. 90° from the compact C_{2v} form and the heavy-atom framework is slightly twisted and its symmetry is also not precisely C_{2v}.

Structural studies of large molecules may be illustrated by cyclopentadienyl chromium dicarbonyl nitrosyl $(\eta^5{-}C_5H_5)Cr(CO)_2NO$, a simple piano-stool-type configuration. Interest in these metal–nitrosyl compounds arises since NO is a neurotransmitter. A study of various isotomers gives the overall molecular configuration and structure of Fig. 15.

2. Free Radicals and Ions

Microwave techniques have been applied to the study of free radicals such as CH, CN, OH, NO, SO (see Table III), ClO, SiF_2, SiN, CCH, HO_2, CuO, and other short-lived species. *In situ* synthesis in the absorption cell is often used to produce these unstable species (see Section XI). The free radicals OH, CH, and CCH have also been identified in interstellar space. Molecules with electronic angular momentum (spin or orbital) have rather complicated rotational spectra that arise because of the interaction of the electronic angular momentum with the rotational angular momentum. The details of the spectrum depend on



r(Cr-Cp)	1.86
r(C1-C2)	1.41
r(Cr-N)	1.69
r(Cr-CO)	1.88
∠Cp-Cr-N	128.
∠Cp-Cr-CO	120.1
∠Cr-N-O	176
r^*(N-O)	1.169
r^*(C-O)	1.145
∠OC-Cr-CO	92.0

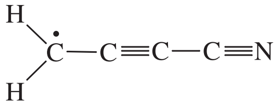
FIGURE 15 A view of η^5-C_5H_5 $Cr(CO_2)_2NO$. Some structural parameters are also given. Cp denotes η^5-C_5H_5; asterisk indicates assumed value. For details see Kukolich, S. G., McKay, R. T., Breckenridge, S. M., Flores, L. D., Morris, G. E., Sèckafoose, S. M., and Morrison, D. L. (1995). *Inorg. Chem.* **34,** 4182.

the type of coupling, and various coupling cases are possible depending on the origin of the electronic angular momentum. The stable O_2 molecule in a $^3\Sigma$ electronic ground state is an important example that has been extensively studied. It has a spin angular momentum of $1\hbar$ and a zero electronic orbital angular momentum. O_2 does not have an electric dipole moment, but does have a magnetic moment associated with the spin angular moment. The corresponding magnetic dipole transitions are, however, much weaker in general than electric dipole transitions.

Free radicals, like ions, play an important role in various chemical reactions. The understanding of chemical mechanisms, whether associated with combustion, the interstellar medium, or other areas, can be clarified by detection of radical intermediates.

Some recent studies of free radicals are given in Table XI. Here, we give the ground electronic state, pertinent comments, and structural parameters for some of the radicals. The reader is directed to the references of the table for a more detailed discussion of the experimental methods and analysis of the rotational spectrum. A number of interesting radicals are given in the table, such as cyclic C_3H, C_nO, C_3H, C_nH, and HC_3O. Also, we note that the shorter OH distance in MgOH as compared to CaOH and other alkaline earth hydroxide radicals indicates the quasilinear nature of this molecule.

The carbon chain radical cyanopropynyl



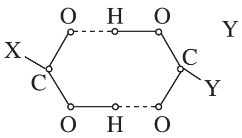
has recently been detected in the laboratory, and the derived rotational constants (in MHz) $B = 2186.430$ and $D = 139 \times 10^6$ provide the basis for astronomical detection. Similarly, observation has recently added C_5N to the radical series CN and C_3N, which have been detected in interstellar space. An initial radio astronomy search for C_5N based on (in MHz) $B = 1403.080$ and $D = 50 \times 10^6$ failed to find lines from this radical.

The observation of molecular ions such as CO^+, HCO^+ (see Table III), HOC^+, HNN^+, NO^+, and HCS^+ represents a relatively new area of study by microwave spectroscopy. These ions have also been detected in outer space by microwave astronomers. Molecular ions represent a significant challenge to microwave spectroscopists, as attested to by the small number of ions so far analyzed. The major problem is in the production of a sufficient concentration of ions for detection (see Section XI). The chemistry produced by molecular ions is important to a number of areas such as plasmas, flames, the upper atmosphere, and the interstellar medium. In the latter case, they play a role in the synthesis of the many neutral molecules so far identified in interstellar space.

Some other molecular ions that have been studied more recently are listed in Table XII. Some examples of ionic complexes are also included in the table. In some cases, the analysis of the rotational spectrum was aided by the use of double resonance techniques or by information obtained from *ab initio* calculations. Most ions have been positive species (cation). Recently, the negative ion (anion) SH^- has been identified in the laboratory. The ion was produced by the low-pressure electric discharge of H_2S and argon. The measurement of this and other light hydrides has been aided by the availability of higher frequency microwave sources. The first two transitions of SH^- fall at 0.56 and 1.1 THz (see Section XI).

3. H-Bonded Complexes and van der Waals Complexes

Microwave spectroscopy studies have provided considerable information on the phenomenon of hydrogen bonding. Earlier studies with conventional microwave spectrometers were on large H-bonded complexes. Typical of these studies is the bimolecule trifluoroacetic acid–cyclopropane carboxylic acid,



$Y = C_3H_5$

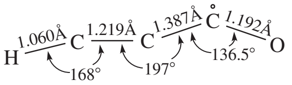
$X = CF_3$ $\quad$ $O\cdots O = 2.62$ Å,

which has recently been studied in the 1- to 4-Ghz region, where the low-J transitions of this heavy complex appear.

A number of hydrogen-bonded complexes that form linear, symmetric, and asymmetric rotors have been studied. Details have been obtained on the geometry, structural parameters, strength of the intermolecular hydrogen bonds, dipole moment, and electric charge distribution of the complexes. This information comes from measurements of rotational constants, centrifugal distortion constants, electric dipole moments, and nuclear hyperfine structure. Tunneling effects and large vibrational motions for such complexes can complicate the analysis. By way of illustration, Tables III and IV give the structures of a few selected complexes. Note the result of $OC\cdots HF$ is consistent with the polarity, O^+C^-.

In a similar way, an ever-expanding body of knowledge is being obtained on inert-atom–molecule complexes. These are weakly bonded van der Waals complexes involving the rare gas atoms Ne, Ar, Kr, and Xe. The formation of weakly polar heteronuclear rare gas dimers allow the pure rotational spectra to be observed. In $Ne\cdots Ar$,

TABLE XI Selected Molecular Radical Studies[a]

Radical	Electronic ground state	Information/comments	Reference
H_2NO	$^2B_1$	Spectroscopic constants, essentially planar, C_{2v}, $r_0(NH) = 1.01$ (assumed), $r_0(NO) = 1.280$, $\angle HNH = 122.7$	Mikami, H., Saito, S., and Yamamoto, S. (1991). *J. Chem. Phys.* **94,** 3415
CH_2N	$^2B_2$	Spectroscopic constants, planar, $r(CH) = 1.08$ (assumed), $r(CN) = 1.261$, $\angle HCH = 122.3$	Yamamoto, S., and Saito, S. (1992). *J. Chem. Phys.* **96,** 4157
c-C_3H	$^2B_2$	Spectroscopic constants, C_{2v}, planar cyclic triangular structure, $r_s(CH) = 1.0760$, $r_s(CC) = 1.3771$, $r_s(C{-}CH) = 1.3739$	Yamamoto, S., and Saito, S. (1994). *J. Chem. Phys.* **101,** 5484
HCCCO	$^2A'$	Spectroscopic constants, planar, 	Cooksy, A. L., Watson, J. K. G., Gottlieb, C. A., and Thaddeus, P. (1994). *J. Chem. Phys.* **101,** 178
MgOH	$^2\Sigma^+$	Spectroscopic constants, linear, $r_0(MgO) = 1.780$, $r_0(OH) = 0.871$[b]	Nuccio, B. P., Apponi, A. J., and Ziurys, L. M. (1995). *J. Chem. Phys.* **103,** 9193
CaOH	$^2\Sigma^+$	Spectroscopic constants, linear, $r_0(CaO) = 1.985$, $r_0(OH) = 0.922$	Nuccio, B. P. *et al.* (1995)
FeCO	$^3\Sigma^-$	Spectroscopic constants, linear, $r_s(FeC) = 1.7268$, $r_s(CO) = 1.1599$	Kasai, Y., Obi, K., Ohshima, Y., Endoy, Y., and Kawaguchi, K. (1995). *J. Chem. Phys.* **103,** 90
C_nO ($n = 2, 4, 6, 8$)	$^3\Sigma^-$	Spectroscopic constants, linear, effective C=C bond[c]: 1.3704, 1.2920, 1.2830, 1.2802 for $n = 2,4,6,8$, respectively	Ohshima, Y., Endo, Y., and Ogata, T. (1995). *J. Chem. Phys.* **102,** 1493
C_3H (C_1,C_2C_3H)	$^2\Pi_r$	Spectroscopic constants, linear,[d] $r_s(C_1C_2) = 1.3263$, $r_s(C_3H) = 1.0171$, $r_s(C_2C_3) = 1.2539$	Kanada, M., Yamamoto, S., Saito, S., and Osamura, Y. (1996). *J. Chem. Phys.* **104,** 2192
CaSH	$^2A'$	Spectroscopic constants, bent, $r_0(CaS) = 2.564$, $r_0(SH) = 1.357$, $\angle CaSH = 91.0$	Taleb-Bendiab, A., Scappini, F., Amano, T., and Watson, J. K. G. (1996). *J. Chem. Phys.* **104,** 7431
FS_2	$^2A''$	Spectroscopic constants, bent, $r_0(FS) = 1.595$, $r_0(SS) = 1.915$, $\angle FSS = 107.8$	Tang, J., and Saito, S. (1996). *J. Chem. Phys.* **104,** 7437; see also Zhuo, Q., *et al.* (1994). *J. Chem. Phys.* **100,** 6113
H_2PO	$^2A'$	Spectroscopic constants, pyramidal, $r_0(PO) = 1.4875$, $r_0(PH) = 1.4287$, $\angle HPO = 115.5$, $\angle HPH = 102.6$	Hirao, T., Saito, S., and Ozeki, H. (1996). *J. Chem. Phys.* **105,** 3450
AsH	$^3\Sigma^-$	Spectroscopic constants, $r_e = 1.52237$	Fujiwara, H., Kobayashi, K., Ozeki, H., Saito, S., and Jaman, A. I. (1997). *J. Chem. Soc. Faraday Trans.* **93,** 1045
FeF	$^6\Delta_i$	Spectroscopic constants, $r_e(FeF) = 1.7803$	Allen, M. D., and Ziurys, L. M. (1997). *J. Chem. Phys.* **106,** 3494
MgBr	$^2\Sigma^+$	Spectroscopic constants, $r_e(MgBr) = 2.3474$	Walker, K. A., and Gerry, M. C. L. (1997). *J. Chem. Phys.* **107,** 9835
NH_2	$^2B_1$	Spectroscopic constants, bent, $r_e(NH) = 1.0254$, $\angle HNH = 102.85$	Kobayashi, K., Ozeki, H., Saito, S., Tonooka, M., and Yamamoto, S. (1997). *J. Chem. Phys.* **107,** 9289
HSiO	$^2A'$	Spectroscopic constants, bent, $\langle r \rangle$: $r(SiO) = 1.5326$, $r(SiH) = 1.5066$ (assumed) $\angle HSiO = 116.8$	Izuha, M., Yamamoto, S., and Saito, S. (1997). *J. Mol. Struct.* **413,** 527
PH	$^3\Sigma^-$	Spectroscopic constants, $\langle r \rangle = 1.43283$	Klisch, E., Klien, H., Winnewisser, G., and Herbst, E. (1998). *Z. Naturforsch.* **53a,** 733
C_nH ($n = 10$, 12,13,14)	$^2\Pi$	Spectroscopic constants, linear carbon chain	Gottlieb, C. A., McCarthy, M. C., Travers, M. J., Grabow, J.-U., and Thaddeus, P. (1998). *J. Chem. Phys.* **109,** 5433

[a] Bond lengths in angstrom units, bond angles $\angle ABC$ in degrees.

Spectroscopic constants: Rotational constants, distortion constants, angular momentum coupling fine structure terms, hyperfine coupling constants, etc. Not all constants are reported for a given molecule. See also references cited.

[b] Short bond may indicate quasilinear character.

[c] Calculated assuming all C=C bonds in the molecule are equal.

[d] Short C—H bond is evidence for quasilinear character.

TABLE XII Selected Molecular Ion Studies[a]

Ion	Information/comments	Reference
SH^-	Spectroscopic constants; method to distinguish between positively and negatively charged ions and neutrals in plasma	Civis, S., Walters, A., Yu Tretyakov, Yu., Bailleux, S., and Bogey, M. (1998). *J. Chem. Phys.* **108,** 8369
HN_2^+	Spectroscopic constants, quadrupole coupling, infrared-microwave double resonance	Ho, W. C., Pursell, C. J., Weliky, D. P., Takayi, K., and Oka, T. (1990). *J. Chem. Phys.* **93,** 87
PO^+	Spectroscopic constants, $r_e = 1.424993$ Å, $\omega_e = 1411.5\ cm^{-1}$, *ab initio* calculations, comparison to SiF^+	Petrmichl, R. H., Peterson, K. A., and Woods, R. C. (1991). *J. Chem. Phys.* **94,** 3504
XeH^+	Spectroscopic constants, $r_e = 1.6028$ Å, $\omega_e = 2270.2\ cm^{-1}$, quadruple coupling, Zeeman analysis, spin–rotation coupling	Peterson, K. A., Petrmichl, R. H., McClain, R. L., and Woods, R. C. (1991). *J. Chem. Phys.* **95,** 2352
$He\cdots Ar^+$	Zeeman analysis, microwave–microwave double-resonance study	Carrington, A., *et al.* (1995). *J. Chem. Phys.* **102,** 2379
FN_2^+	Spectroscopic constants, *ab initio* calculations, r(F—N) = 1.2461 Å, r(N—N) = 1.1034 Å	Botschwina, P., Sebald, P., Bogey, M., Demuynck, C., and Destombes, J.-L. (1992). *J. Mol. Spectrosc.* **153,** 255
FCO^+	Spectroscopic constants, *ab initio* calculations, r(F—C) = 1.2014 Å, r(C—O) = 1.1151 Å	Botschwna, P., *et al.* (1992)
$Ar\cdots H_3^+$	Spectroscopic constants, large-amplitude internal rotation motion, 120° planar rotation of the H_3^+ group relative to Ar atom; $r = 2.3852$ Å is the distance from the center of mass of the H_3^+ triangle to Ar	Baillex, S., *et al.* (1998). *J. Mol. Spectrosc.* **190,** 130
$HNCCN^+$	Spectroscopic constants, linear configuration	Amano, T., and Scappini, F. (1991). *J. Chem. Phys.* **95,** 2280

[a] See also references cited. Spectroscopic contants: Rotational constants, distortion constants, various other interaction constants.

for example, the small induced dipole is 0.0022 D. These and other weakly bonded complexes are usually produced by the rapid expansion of a pressurized mixture of the constituents through a supersonic nozzle into an evacuated cell of the spectrometer (see Section XI). Information like that outlined above for H-bonded complexes can also be attained for these complexes. These studies of weakly bonded complexes provide important information about intermolecular forces. Recently, the range of complexes studied has been extended to include clusters such as $(Ar)_n$—HF ($n = 2,3,4$) and $(CO_2)n$—HF ($n = 2,3$).

In the analysis of complexes, the accurate structures, barriers to internal rotation, etc., of the molecular species involved in complex formation are important in order to understand the subtle changes that may occur upon complexation. Also, it may be noted that for a complex formed by an asymmetric-top molecule and a rare gas atom, there are, in general, eight structures which are compatible with the moments of inertia. This ambiguity can often be clarified or reduced by use of symmetry arguments or other parameters such as the dipole moment, nuclear quadrupole structure, isotopic information, or *ab initio* calculations. The latter has been particularly useful in clarifying structural ambiguities and providing insight into the most stable form. A large number of weakly bound complexes have been studied by microwave spectroscopy.

Table III gives a few molecular structures. In addition, Table XIII provides the geometries of some selected complexes.

4. Rotational Isomerism and Ring Conformations

Microwave spectroscopy has also provided considerable information on rotational isomerism and conformations of ring compounds. The existence of one or more distinct rotational isomers for a large number of molecules has been demonstrated. Similarly, the stable conformations for numerous ring compounds have been derived. Detailed structural information has also been obtained for many rotamers and ring systems.

The sensitivity of the rotational constants to the ring conformation is illustrated in Fig. 16. The boat conformation is clearly the correct conformation. Similarly, the rotational isomers evident in Fig. 10 may be identified by comparison of the observed and calculated separation ($B + C$) between the bands of a given series.

Distortion of small rings due to substitution has also been observed. In going from cyclopropene, $\overline{CH{=}CHCH_2}$, to 3,3-difluorocyclopropene, $\overline{CH{=}CHCF_2}$, for example, the C=C double bond is increased in length by 0.025 Å, while the other ring bonds are shortened. On the other hand, axiridine, $\overline{CH_2CH_2NH}$, and chloroaziridine, $\overline{CH_2CH_2NCl}$, are remarkable in that detailed structural studies reveal that the ring structures are virtually identical. Isoxazole, $\overline{OCH{=}CHCH{=}N}$, is a good example of the tendency of C—H bonds adjacent to ring heteroatoms to tilt slightly toward these atoms.

A detailed study of various isotopomers of 1,2,3-trifluorobenzene showed that the bond angles at C_1 and C_3 are enlarged, while the C_1—C_2 and C_2—C_3 bonds are

TABLE XIII Geometries for Selected Complexes

Complex	Geometry	Reference
	Triangular	Bumgarner, R. E., and Kukolich, S. G. (1987). *J. Chem. Phys.* **86,** 1083
	Planar, T-shaped, C_{2V}	Legon, A. C., Aldrich, P. D., and Flygare, W. H. (1981). *J. Chem. Phys.* **75,** 625
	Cis, planar	Kukolich, S. G. (1983). *J. Mol. Spectrosc.* **98,** 80
	Nonplanar, cyclic	Bumgarner, R. E., Pauley, D. J., and Kukolich, S. G. (1987). *J. Chem. Phys.* **87,** 3749
	Planar, T-shaped, C_{2V}	Gutowsky, H. S., Klots, T. D., Chuang, C., Schmuttenmaer, C. A., and Emilsson, T. (1987). *J. Chem. Phys.* **86,** 569
	C_{3V} symmetric top, H—F along C_3 axis of Ar_3 group, H closest to Ar_3	Gutowsky, H. S., Klots, T. D., Chuang, C., Keen, J. D., Schmuttenmaer, C. A., and Emilsson, T. (1987). *J. Am. Chem. Soc.* **109,** 5633
	Symmetric top, HCN along C_3 axis of cyclic $(CO_2)_3$ group, N closest to $(CO_2)_3$	Gutowsky, H. S., Hajduk, P. J., Chuang, C., and Ruoff, R. S. (1990). *J. Chem. Phys.* **92,** 862
	Linear HCN trimer	Ruoff, R. S., Emilsson, T., Klots, T. D., Chuang, C., and Gutowsky, H. S. (1988). *J. Chem. Phys.* **89,** 138
	T-shaped, C_{2v} symmetry	Lida, M., Ohshima, Y., and Endo, Y. (1991). *J. Chem. Phys.* **94,** 6989
	Ar above thiazole ring shifted toward N atom	Kretschmer, U., Stahl, W., and Dreizler, H. (1995). *J. Mol. Struct.* **352,** 289
	T-shaped, C_{2v} symmetry	Mader, H., Heineking, N., Stahl, W., Jager, W., and Xu, Y. (1996). *J. Chem. Soc. Faraday Trans.* **92,** 901
	S···Cl—Cl collinear, Cl_2 nearly perpendicular to H_2S plane	Bloemink, H. I., Dolling, S. J., Hinds, K., and Legon, A. C. (1995). *J. Chem. Soc. Faraday Trans.* **91,** 2059
	Water-donor, methanol-acceptor complex	Stockman, P. A., Blake, G. A., Lovas, F. J., and Suenram, R. D. (1997). *J. Chem. Phys.* **107,** 3782
	Induced dipole ~0.01 D	Jager, W., Xu, Y., and Gerry, M. C. L. (1993). *J. Chem. Phys.* **99,** 919
	Heavy atoms, approximately collinear, H_2O···O hydrogen bond	Yaron, D., Peterson, K. I., Zolandz, D., Klemperer, W., Lovas, F. J., and Suenram, R. D. (1990). *J. Chem. Phys.* **92,** 7095
	Xe, Kr complexes T-shaped, Ne complex tilted as shown	Walker, K. A., Ogata, T., Jager, W., Gerry, M. C. L., and Ozier, I. (1997). *J. Chem. Phys.* **106,** 7519

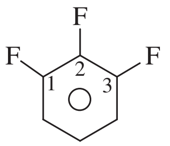
continues

TABLE XIII (*Continued*)

	Ar above furan plane shifted toward O with respect to furan center of mass	Spycher, R. M., Hausherr-Primo, L., Grassi, G., and Bauder, A. (1995). *J. Mol. Struct.* **351,** 7
	C_{2v} symmetry, triangular trimer	Xu, Y., Jager, W., and Gerry, M. C. L. (1994). *J. Chem. Phys.* **100,** 4171
	C_s symmetry, with terminal oxygens of O_3 tilted toward axis of acetylene	Gillies, J. Z., Gillies, C. W., Lovas, F. J., Matsumura, K., Suenram, R. D., Kraka, E., and Cremer, D. (1991). *J. Am. Chem. Soc.* **113,** 6408

[a] See also Novick, S. E., Leopold, K. R., and Klemperer, W. (1990). "Atomic and molecular clusters." *In* "The Structures of Weakly Bond Complexes As Elucidated by Microwave and Infrared Spectroscopy" (E. R. Bernstein, ed.), Chapter 3, Elsevier, New York.

slightly shortened compared to the opposite end of the benzene ring:



Small changes in structural parameters in going from one rotamer to another may be illustrated by *cis-* and *trans-*

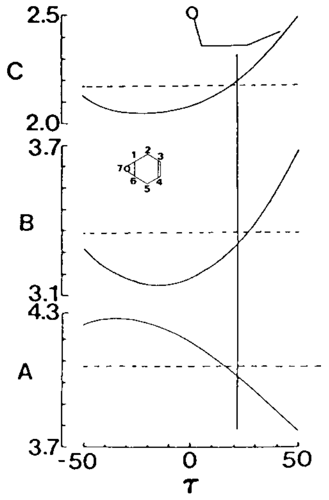
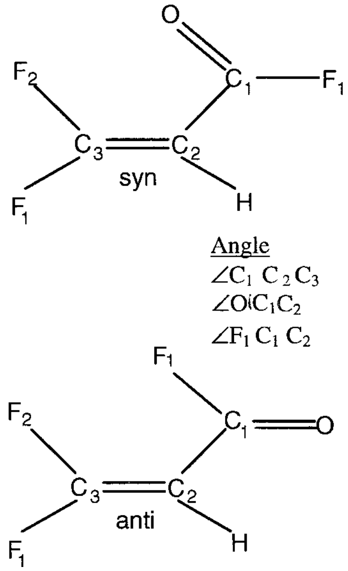


FIGURE 16 Calculated rotational constants of 7-oxabicyclo [4.1.0] hept-3-ene for various values of the angle τ. τ measures the displacement of atoms 3 and 4 out of the plane defined by atoms 1, 2, 5, and 6. Positive values of τ correspond to the boat form, negative values to the chair form. Observed rotational constants (*A*, *B*, and *C*) are indicated by horizontal broken lines. The units of the rotational constants are GHz. The good agreement between observed and calculated rotational constants for the boat conformation established this as the correct form.

monothioformic acid, HCOSH. In going from the *trans* to the *cis* rotamer, the CSH and the HCS angles increase by more than 2°, whereas the OCS angle decreases by 3.4°. The SH bond length shortens by 0.02 Å.

Similarly, subtle changes in structure which occur in different isomers are illustrated for 3,3-difluoroacryloyl fluoride in Fig. 17, where the $C_3C_2C_1$ and $C_2C_1F_1$ angles increase, while C_2C_1O decreases in going from the *syn* to the *anti* conformer.

Microwave spectroscopy has also provided direct evidence as to the importance of intramolecular hydrogen



Angle	*Anti*	*Syn*
$\angle C_1C_2C_3$	126.6°	123.8°
$\angle OC_1C_2$	124.3°	130.9°
$\angle F_1C_1C_2$	115.2°	108.7°

FIGURE 17 Conformers of difluoroacryloyl fluoride and structural changes on *syn–anti* isomerization. For details of the analysis see Tam, H. S., Harmony, M. D., Brahms, J. C., and Dailey, W. P. (1991). *J. Mol. Struct.* **244,** 59.

bonding in determining the conformations (often the more stable one) of certain rotamers. For methoxyacetic acid, CH_3OCH_2COOH, for example, an intramolecular hydrogen bond is formed between the hydroxyl hydrogen and the ether oxygen, giving a planar five-membered ring.

5. Methyl Group: Tilts and Asymmetry

Numerous examples of methyl group tilts have been found where the symmetry axis of the methyl group does not correspond with the bond direction. The effect has been demonstrated from analysis of internal rotation splittings and directly from structure determinations. In methyl nitrate, CH_3ONO_2, for example, the CH_3 groups is tilted toward the unshared electron pair by 4.8°. The NO_2 group is also found to be tilted by 3.9° away from the methyl group.

Asymmetric methyl groups for which the hydrogen atom in-plane (plane of the heavy-atom skeleton) has a slightly smaller C—H bond length than the out-of-plane hydrogens have been reported. Examples are methyl nitrate and thioacetaldehyde.

6. Quasi-Linear Molecules

Finally, we mention HCNO, where the ground-state spectrum is consistent with a linear structure, but the equilibrium structure is slightly bent. It has been found that the bending potential function has a small hump well below the ground vibrational level at the linear configuration. HCNO is one of a handful of molecules, called quasi-linear molecules, with this characteristic behavior. Other examples are HNCO, HNCS, and HNCSe. In a similar way, CH_3NCO, CH_3NCS, and CH_3NCSe are examples of quasi-symmetric-top molecules.

7. Interstellar Space Applications

The region between the stars contains interstellar clouds composed primarily of dust and gas. Over 100 molecules (neutrals, ions, and radicals), mostly carbon-containing compounds, have been identified in interstellar molecular clouds. The largest molecule contains 13 atoms, the carbon chain $HC_{11}N$.

Interstellar spectroscopy is usually carried out with a single-dish radio telescope composed of a precision parabolic reflector with a highly sensitive microwave detector at the focal point of the reflector. Molecular identifications, usually via rotational emission frequencies, are made by comparison with precise laboratory frequency measurements or frequency predictions. Microwave spectroscopy plays a very important role in providing the data for identification of new molecular species. Identifications have been made mostly with data in the microwave, millimeter-, and submillimeter-wave regions. Active interplay between laboratory observations and astronomical observations has helped to drive both areas.

Laboratory measurements have been driven by continued improvements in experimental techniques such as rapid cooling Fourier transform microwave spectroscopy (see Section XI). This technique, coupled with improved species production methods, has provided capabilities to identify an increasing number of exotic molecular species.

Moreover, recent advances in methods to generate high-frequency microwave radiation, such as high-frequency backward wave oscillators, have enabled laboratory measurements in the terahertz region (1 THz = 1000 GHz, equivalent to $\lambda = 0.3$ mm). This technological development will enable, for example, the direct detection of light hybrids for which even the lowest rotational transitions are at high frequencies. In the case of the radical NH, for example, with a $^3\Sigma$ electronic ground state, the ground-state transition is split into three fine-structure transitions at frequencies 946, 974, and 1000 GHz. Each of these transitions is split by quadruple (^{14}N) and magnetic hyperfine (H) interactions.

The detection of molecules via their rotational spectra allows astrophysicists to probe interstellar clouds to provide information on their environment, star formation, interstellar chemistry, mechanisms for synthesis and destruction of interstellar molecules, isotopic distributions, etc.

Continued and increased advances in our knowledge of interstellar space can be expected. The satellite FIRST (Far Infrared and Submillimeter Space Telescope) is planned to be launched by 2003 and will cover the frequency range 300–3000 GHz ($\lambda = 1$–0.1 mm). This telescope will also provide information on planets and comets. These satellite observations, coupled with high-frequency microwave studies, will enable the identification of a host of other molecules in space.

V. CENTRIFUGAL DISTORTION EFFECTS

The rigid-rotor treatment discussed in the previous section accounts for the general features of the rotational spectrum. These gross features are modified somewhat when the effects of nonrigidity, nuclear coupling, and so forth are taken into account. In this section, the effects of centrifugal distortion are considered.

The centrifugal force produced by rotation distorts the molecule from its equilibrium configuration and the bond distance and angles change slightly. Hence, the rotational spectrum is no longer characterized by a set of equilibrium moments of inertia. Additional terms in the Hamilitonian are required to account adequately for the observed

spectrum. If distortion is taken into account, terms in the angular momentum components of the fourth power, sixth power, and so forth are introduced into the Hamiltonian. Although the centrifugal distortion constant are very small relative to the rotational constants, they produce significant effects in the rotational spectrum, particularly for asymmetric tops.

A. Diatomic and Linear Rotors

For diatomic and linear rotors, distortion terms of the type DP^4 and HP^6 are added to the rigid-rotor Hamilitonian. Considering only the major P^4 effect and noting $(J, M|P^4|J, M) = \hbar^4 J^2(J+1)^2$, we have for the rotational energy of a nonrigid diatomic or linear molecule

$$E_J = BJ(J+1) - D_J J^2(J+1)^2 \quad (50)$$

and the frequencies are given by

$$\nu = 2B(J+1) - 4D_J(J+1)^3, \quad (51)$$

where D_J is the distortion constant, and both B and D_J are in frequency units. The physical picture of distortion in a diatomic molecule is quite simple. As the molecular bond stretches, I increases and B decreases, leading to a decrease in the rotational energy and a shift to lower frequency relative to the rigid rotor frequency. If P^6 effects are considered, then a term $H_J J^3(J+1)^3$ must be added to Eq. (50). Distortion effects are small and primarily important only for high J values. Some typical values of distortion constants are listed in Table III.

B. Symmetric-Top Molecules

A first-order treatment of centrifugal distortion yields for the distortion Hamiltonian

$$\mathcal{H}_d = -D_J P^4 - D_{JK} P^2 P_Z^2 - D_K P_Z^4, \quad (52)$$

which is diagonal in the symmetric-top basis since the only nonvanishing matrix elements (in units of $\hbar$) are $(J, K, M|P^4|J, K, M) = J^2(J+1)^2$, $(J, K, M|P^2 P_Z^2|J, K, M) = K^2 J(J+1)$, and $(J, K, M|P_Z^4|J, K, M) = K^4$. The energy of a nonrigid prolate symmetric top is

$$E_{J,K} = BJ(J+1) + (A-B)K^2 - D_J J^2(J+1)^2 - D_{JK} J(J+1)K^2 - D_K K^4. \quad (53)$$

For an oblate symmetric top, the unique axis is designated c, and the energy expression may be obtained from the above expression by replacement of A by C. The constants D_J, D_K, and D_{JK} essentially represent the distortion effects of end-over-end rotation, rotation about the symmetry axis, and the interaction between these motions, respectively. With the selection rules $J \to J+1$, $K \to K$, the rotational frequencies are found to be

$$\nu = 2B(J+1) - 4D_J(J+1)^3 - 2D_{JK}(J+1)K^2. \quad (54)$$

Note that neither A nor D_K affects the rotational spectrum. The first correction term, involving D_J, alters the even spacing between successive $J \to J+1$ transitions, while the last term also separates the superposed lines of different K values into $J+1$ closely spaced lines with the separation increasing as K^2. This is illustrated in Fig. 7. Typical values of D_J and D_{JK} are collected in Table IV. D_J is always positive, whereas D_{JK} and D_K may be positive or negative.

Higher order effects introduce additional distortion terms and also distortion terms that can give rise to splittings of certain K-levels. The effects of centrifugal distortion on the observation of forbidden $\Delta K = \pm 3$ transitions have already been mentioned. Induced dipole moments also allow the observation of pure rotational spectra of spherical tops which, because they have no permanent dipole moment, would otherwise have no rotational spectra. For CH_4, the distortion moment is on the order of 5×10^{-6} D. Both $J \to J+1$ and $J \to J$ transitions have been observed. The leading terms in the frequency equation for the $J \to J+1$ transitions are like those for a linear molecule, Eq. (51); however, the molecular distortion in such molecules is more complicated, and additional terms are required to adequately characterize the rotational spectrum. Such observations have provided the rotation and distortion constants. Some examples of nonpolar molecular studies via microwave spectroscopy are spherical tops with T_d symmetry like CH_4, SiH_4, and GeH_4 and those with D_{3h} symmetry like BF_3 and SO_3. For SO_3, a planar molecule, the centrifugally induced rotational spectrum provides $r_e = 1.4175$ Å.

C. Asymmetric-Top Molecules

The evaluation of centrifugal distortion in asymmetric rotors is considerably more complex than for linear or symmetric tops, and because of the nature of the spectrum, particularly large distortion shifts (say 1000 MHz or larger) can be observed. A first-order treatment of the P^4 distortion effects gives, for the energy of a semirigid prolate asymmetric rotor ($Z \leftrightarrow a$),

$$E = E_r + E_d \quad (55)$$

$$E_r = (1/2)(B+C)J(J+1) + [A - (1/2)(B+C)]W(b_p) \quad (56)$$

$$E_d = -\Delta_J J^2(J+1)^2 - \Delta_{JK} J(J+1)\langle P_Z^2\rangle - \Delta_K \langle P_Z^4\rangle - 2\delta_J \sigma J(J+1)\left[W(b_p) - \langle P_Z^2\rangle\right] - 2\delta_K \sigma\left[W(b_p)\langle P_Z^2\rangle - \langle P_Z^4\rangle\right], \quad (57)$$

where σ is $-1/b_p = (2A - B - C)/(B - C)$, $W(b_p)$ is the Wang reduced energy, and $\langle P_Z^n \rangle$ is the average of P_Z^n in the rigid asymmetric rotor basis $|J, \tau, M)$. These latter quantities may be calculated, as mentioned previously, from the eigenvectors obtained in diagonalization of the rigid-rotor energy matrix. For very slightly asymmetric tops, $\langle P_Z^n \rangle \cong K^n$. The A, B, and C are effective rotational constants that now contain a small contribution involving the distortion constants. The Δ_J, Δ_{JK}, and so on are the quartic distortion coefficients. One may apply the above expressions to an oblate top ($Z \leftrightarrow c$) by interchanging A and C and setting $\sigma = -1/b_o$ in the above energy expression. To evaluate the rotational and distortion constants, differences between the observed and calculated rigid-rotor frequencies are analyzed via Eq. (55) by means of the least-squares technique to determine adjustments in the original rotational constants δA, δB, and δC as well as the distortion constants Δ_J and so on. Quartic distortion constants for a few asymmetric tops are collected in Table V.

The energy expression given in Eq. (57) can account for a large number of asymmetric tops. However, for light molecules with large rotational energies, such as H_2O, or when transitions from high J levels are studied, a first-order treatment does not suffice. Additional higher power terms in the angular momentum must be included in Eq. (57). Specifically, it is found that $n+1$ distortion contributions are added for each degree n in the angular momentum. Thus seven terms are added if P^6 effects are considered. The sextic distortion constants are denoted by Φ_J, Φ_{JK}, Φ_{KJ}, Φ_K, ϕ_J, ϕ_{JK}, and ϕ_K and have been evaluated from a study of the rotational spectra for a large number of molecules. When such effects are important, a first-order treatment is not sufficient. In such cases, the energy matrix of $\mathscr{H}_r + \mathscr{H}_d$ must be set up and diagonalized to obtain the general effects of centrifugal distortion. Procedures for effectively including such higher order distortion effects are discussed elsewhere.

D. Information from Distortion Constants

The study of centrifugal distortion provides a number of useful kinds of information. By including effects of centrifugal distortion, one can obtain very accurate spectroscopic constants. These allow the prediction of unmeasured transition frequencies with a high degree of confidence over a wider range than provided by only the rigid-rotor constants. Most importantly, however, the centrifugal distortion constants provide information on the vibrational potential function, particularly for small molecules. This follows because the distortion constants depend directly on the force constants, masses, and structure of the molecule. This may be illustrated for a diatomic molecule. The quadratic potential function is given by

$$V = (1/2) f R^2, \tag{58}$$

where f is the stretching force constant and R represents the displacement coordinate, which measures the departure of the bond length from its equilibrium value. The constant D_J is defined by

$$D_J = \left(\frac{\hbar^4}{h}\right) \frac{m}{f(I_e)^3} = \frac{4B_e^3}{\omega_e^2}, \tag{59}$$

with m the reduced mass; I_e and B_e are, respectively, the equilibrium moment of inertia and rotational constant $B_e = h/8\pi^2 I_e$. Here $\omega_e = (1/2\pi)(f/m)^{1/2}$ is the harmonic vibrational frequency. Thus from an analysis of the rotational spectrum, precise values of D_J can be obtained, which in turn yield, from the above expression, accurate stretching force constants, or equivalently, ω_e values.

For other molecules, the details are more complicated, but the principles are the same. The quadratic potential function has the general form

$$V = \frac{1}{2} \sum f_{ij} R_i R_j. \tag{60}$$

Infrared measurements yield the vibrational frequencies associated with the various normal vibrational modes, and these data, including isotopic frequency data, can be used to evaluate the force constant matrix $\mathbf{F} = [f_{ij}]$. Since these calculations are often ill conditioned and also since there are usually more force constants than vibrational frequency data, both infrared data and the microwave distortion constant data are often combined to help characterize the force constant matrix. Some examples are given in Table XIV. It may be noted that the P^6 or sextic distortion constants depend on the cubic potential constants, and these data have been employed to obtain information on these anharmonic potential constants.

TABLE XIV Potential Constants Determined by Combination of Infrared and Microwave Data (mdyn/A)[a]

Molecule	f_r	f_{rr}	$f_{\alpha r^2}$	$f_{r\alpha r}$
SO_2	10.006	0.024	0.793	0.189
O_3	5.70	1.52	1.28	0.332
OF_2	3.950	0.806	0.724	0.137
ClO_2	7.018	−0.170	0.651	0.006
NO_2	11.043	2.140	1.109	0.481
GeF_2	4.08	0.26	0.316	−0.01
SeO_2	6.91	0.03	0.488	0.009

[a] The quadratic-valence force field potential function for bent triatomic molecules XY_2 is defined by

$$2V = f_r(\delta r_1^2 + \delta r_2^2) + f_\alpha \delta\alpha^2 + 2 f_{r\alpha}(\delta r_1 + \delta r_2)\delta\alpha + 2 f_{rr}\delta r_1 \delta r_2.$$

VI. ROTATION–VIBRATION INTERACTIONS

In addition to centrifugal distortion effects, other nonrigidity effects also alter the rotational spectra. The vibrational motions of a polyatomic molecule may be described in terms of n normal modes of vibration, where $n = 3N - 6$ (or $3N - 5$ for a linear molecule) with N the number of atoms. For linear and symmetric tops, degenerate vibrations are present, and not all of these modes have different vibrational frequencies ω_i. The modes with the same ω_i are usually grouped together and their number specified by d_i. As a molecule rotates, it vibrates rapidly, even in the ground vibrational state, and the moments of inertia are averaged in a complicated way over the molecular vibrations. The rotational constants as well as the distortion constants, and in fact almost all the molecular parameters derived from an analysis of rotational spectra, must be considered as effective constants, that is, constants associated with a particular vibrational state. The dependence of the effective rotational constants on the vibrational state v may, to a good approximation, be expressed by

$$A_\mathrm{v} = A_\mathrm{e} - \sum \alpha_i^a \left(v_i + \frac{d_i}{2} \right), \tag{61}$$

$$B_\mathrm{v} = B_\mathrm{e} - \sum \alpha_i^b \left(v_i + \frac{d_i}{2} \right), \tag{62}$$

$$C_\mathrm{v} = C_\mathrm{e} - \sum \alpha_i^c \left(v_i + \frac{d_i}{2} \right), \tag{63}$$

where A_e and so on are the equilibrium rotational constants associated with the vibrationless state, v is specified by the vibrational quantum numbers $(v_1, v_2, \ldots, v_i, \ldots)$, where v_i is the quantum number of the ith vibration, and d_i is the corresponding degeneracy. The α_i^a, α_i^b, and α_i^c are the rotation–vibration constants for the ith mode and the A, B, and C rotational constants, respectively. The sum is over the various vibrations, with degenerate vibrations counted only once. For diatomic and asymmetric tops, $d_i = 1$. The dependence on vibrational state of the effective distortion constants is similar to the above expressions, that is, for D_J of a symmetric top, we may write

$$D_J^{(\mathrm{v})} = D_\mathrm{e} + \sum \beta_i (v_i + d_i/2), \tag{64}$$

where β_i is a small rotation–vibration constant. In many cases, particularly in the literature, the rotation or distortion constants are designated simply A or D_J, and the symbol v is omitted. However, it is to be understood that such parameters are, in general, dependent on the vibrational state.

It is apparent from the above expressions that the effective rotational constants as well as the distortion constants are different for each vibrational state, and a separate rotational spectrum is obtained for each vibrational state. These excited-state lines, or satellite spectra, may be shifted only a few megahertz or many hundreds of megahertz away from the ground-state line. Because of the Boltzmann factor $e^{-E_\mathrm{v}/kT}$, the line intensity decreases with increasing vibrational excitation, and only low-lying vibrational levels give rise to lines with sufficient intensity to be observed. Except for diatomic molecules and a few relatively simple polyatomic molecules, it is not possible to obtain sufficient data to determine all the α_i in Eqs. (61)–(63). Hence, it is not possible to correct the observed ground-state rotational constants to obtain the equilibrium constants, for instance,

$$A_\mathrm{e} = A_0 + \sum \alpha_i^a d_i / 2. \tag{65}$$

Because of this, effective constants must be used to evaluate the molecular structure, which introduces uncertainties in the derived structural parameter. This is discussed further in Section VIII.

Unless there is an accidental near-vibrational degeneracy, the rotational spectrum of an asymmetric top in an excited vibrational state is similar to that obtained in the ground state, except that the spectrum is characterized by a slightly different set of rotation and distortion constants. Other nonrigid effects are often more important for asymmetric tops, such as internal rotation, and these are considered in Section VII. Similar statements apply to linear and symmetric-top molecules in excited nondegenerate vibrational states. For example, the rotational frequencies for symmetric tops in nondegenerate vibrational states are given by Eq. (54) with the rotation and distortion constants replaced by effective constants B_v, $D_J^{(\mathrm{v})}$, $D_{JK}^{(\mathrm{v})}$. On the other hand, when degenerate bending modes are present, as with linear and symmetric tops, the spectrum in these excited states can be altered markedly. This effect is called l-type doubling and will be discussed for linear molecules after the general expression for the rotation–vibration energy levels is given for a diatomic molecule.

A. Diatomic Molecules

For diatomic molecules, since there is only one vibrational mode, enough excited states can be studied to enable the evaluation of a number of rotation–vibration constants. With the assumption of a Morse potential, the eigenvalue equation, $\mathcal{H}\psi = E\psi$, for a diatomic molecule can be solved directly, and the energy levels are specified by

$$E_{v,J} = \omega_\mathrm{e}\left(v + \tfrac{1}{2}\right) - \omega_\mathrm{e}\chi_\mathrm{e}\left(v + \tfrac{1}{2}\right)^2 + B_v J(J+1) - D_v J^2(J+1)^2 + H_v J^3(J+1)^3 + \cdots. \tag{66}$$

The first two terms represent the vibrational energy and the last terms the effective rotational energy. The

v (=0, 1, 2, ...) and J (=0, 1, 2, ...) are, respectively, the vibrational and rotational quantum numbers, ω_e is the harmonic vibrational frequency, and $\omega_e\chi_e$ is the anharmonicity constant. The effective rotation and distortion constants are defined by

$$B_v = B_e - \alpha_e\left(v + \tfrac{1}{2}\right) + \gamma_e\left(v + \tfrac{1}{2}\right)^2 + \cdots, \quad (67)$$

$$D_v = D_e + \beta_e\left(v + \tfrac{1}{2}\right) + \cdots, \quad (68)$$

$$H_v = H_e + \cdots. \quad (69)$$

The α_e, γ_e, and β_e are the rotation–vibration interaction constants representing corrections for the effect of vibration. The selection rules for pure rotational transitions are $J \to J + I$, $v \to v$, and the rotational frequencies are easily shown to be

$$\nu = 2B_v(J+1) - 4D_v(J+1)^3 + H_v(J+1)^3 \times [(J+2)^3 - J^3]. \quad (70)$$

To evaluate all of the constants, measurements of rotational transitions in at least three vibrational states (e.g., $v = 0$, 1, and 2) must be made. Each vibrational state is analyzed via the above equation. For example, from the data B_0, B_1, and B_2, Eq. (67) yields

$$B_e = \tfrac{1}{8}(15B_0 - 10B_1 + 3B_2). \quad (71)$$

Some spectroscopic constants obtained for a few selected diatomic molecules are collected in Table XV. Information on the vibrational constants can also be obtained from the rotational constants, for example,

$$\omega_e^2 = \frac{4B_e^3}{D_e}, \quad (72)$$

$$\omega_e\chi_e = B_e\left(\frac{\alpha_e\omega_e}{6B_e^2} + 1\right)^2. \quad (73)$$

B. *l*-Type Doubling in Linear Molecules

For linear molecules, the rotational frequencies in excited nondegenerate vibrational states are specified by Eq. (51). Furthermore, the effects of vibration on B_v and D_v are given by Eqs. (62) and (64). However, the rotational spectrum is complicated by the presence of degenerate bending modes of vibration. The bending mode is twofold degenerate since the linear molecule may bend in either of two orthogonal planes. In the case of OCS, for example, there are $3 \cdot 3 - 5 = 4$ vibrational modes, labeled v_1, v_2, and v_3 with the bending mode twofold degenerate, $d_2 = 2$. With excitation of a single degenerate bending mode v_j, an angular momentum $p = l\hbar$ is generated along the molecular axis analogous to a symmetric top with l similar to K. The possible values of l are

$$l = v_j, v_{j-2}, v_{j-4}, \ldots, -v_j. \quad (74)$$

Hence, for $v_j = 1$, $l = \pm 1$; $v_j = 2$, $l = 0, \pm 2$; and so on. In addition, a Coriolis interaction between rotation and vibration exists that can remove the $\pm l$ degeneracy when $l \neq 0$. The linear molecule behaves in an excited bending state as if it were slightly bent; and, like a slightly asymmetric rotor, where the $\pm K$ degeneracy is lifted, the $\pm l$ degeneracy is lifted. This is called l-type doubling. A detailed treatment for the energies including the l-type splitting of the levels gives

$$E = B_v[J(J+1) - l^2] - D_v[J(J+1) - l^2]^2 \pm \tfrac{1}{4}q_j(v_j + 1)J(J+1), \quad (75)$$

where q_j is the coupling constant characterizing, the splitting for the bending mode v_j. This constant is usually significant only for the case $|l| = 1$. Here B_v and D_v have their usual meaning. If the splitting term is omitted, it is apparent that the levels are doubly degenerate since they depend on l^2, except when $l = 0$. However, J represents the total angular momentum quantum number including the vibrational angular momentum. Hence,

$$J = |l|, |l| + 1, |l| + 2, \ldots, \quad (76)$$

and depending on l, certain values of J are missing. In particular, for $|l| = 1$, $J = 1$ is the lowest value of J, while for $|l| = 2$, $J = 2$ is the lowest value. As a result of this,

TABLE XV Selected Molecular Constants of Some Diatomic Molecules

Diatomic molecule	B_e (MHz)	$-\alpha_e$ (MHz)	D_e (kHz)	ω_e (cm^{-1})	$\omega_e\chi_e$ (cm^{-1})	r_e (Å)
$^{28}Si^{16}O$	21,787.453	151.026	29.38	1252	5.96	1.50973
$^{74}Ge^{32}S$	5,593.1019	22.4569	2.41	569	1.723	2.0120772
$^{74}Ge^{130}Te$	1,958.7903	5.1702	0.353	308	0.62	2.3401556
$^{120}Sn^{16}O$	10,664.189	64.243	7.98	882	3.93	1.832198
$^{120}Sn^{32}S$	4,103.0013	15.1585	1.272	491.6	1.412	2.2090172
$^{208}Pb^{32}S$	3,487.1435	13.0373	1.012	431.8	1.277	2.2868535
$^{208}Pb^{80}Se$	1,516.9358	3.8952	0.210	272.3	0.552	2.402223

for example, the $0 \rightarrow 1$ transition is missing in an excited bending mode with $|l| = 1$, and the $0 \rightarrow 1$ and $1 \rightarrow 2$ transitions are missing for $|l| = 2$. With the selection rules $J \rightarrow J + 1, l \rightarrow l$, the rotational frequencies including l-type doubling are

$$\nu_{\pm} = 2B_v(J+1) - 4D_v(J+1)[(J+1)^2 - l^2] \pm \tfrac{1}{2}q_i(v_j+1)(J+1). \quad (77)$$

If l-type doubling represented by the last term is negligible, the rotational spectrum is like that for a linear molecule in a nondegenerate vibrational state except for the limitations on the values of J and small effects in the distortion correction because of the presence of the vibrational angular momentum quantum number l.

To distinguish the vibrational states, the value of $|l|$ is added as a superscript to the vibrational quantum number associated with the degenerate vibration. For a triatomic linear molecule, the states are specified by $(v_1, v_2^{|l|}, v_3)$. Thus, the notation $(1, 2^2, 0)$ corresponds to a state with the v_1 stretching mode excited by one unit, the v_3 stretching mode in its ground state, and the bending mode v_2 excited by two units with $l = \pm 2$. The nonrigid rotor spectrum of FC≡P is illustrated in Fig. 18. Excitation of the degenerate vibrational mode produces a series of lines of decreasing intensity and l-type doubling is apparent when $l = \pm 1$.

It is also possible to observe direct transitions between the closely spaced l-type doublet levels with $\Delta J = 0$. The l-doublet transitions for $|l| = 1$ are given by

$$\nu = (1/2)q_j(v_j+1)J(J+1). \quad (78)$$

With sufficiently large q_j or J, these transitions can fall into the microwave region. For HCN, $v_2 = 1, q_2 =$ 224.48 MHz, and a series of transitions is found in the microwave region. In practice, it is found that q_j varies slightly with J, and more elaborate expressions are required to fit accurately the l-type doublet spectra.

To define the rotation–vibration constants in terms of more fundamental parameters or to understand the origin of various nonrigidity effects in the rotational spectrum, the general rotation–vibration Hamiltonian must be employed. This Hamiltonian contains pure rotation and vibration terms as well as interaction terms between rotation and vibration. Perturbation treatments to various orders are required to characterize the different rotation–vibration effects. Space does not permit further discussion of this; however, we mention that such a perturbation treatment shows that the α constants depend on the cubic potential energy constants of the molecule.

VII. INTERNAL ROTATION

Internal rotation involves the rotation of one part of a molecule relative to the other about a single bond. The appearance of the rotational spectrum depends on the type of internal rotor and on the barrier height hindering internal rotation. Rotation of a methyl group, hindered by a barrier on the order of 3 kcal/mole, leads to a splitting of the

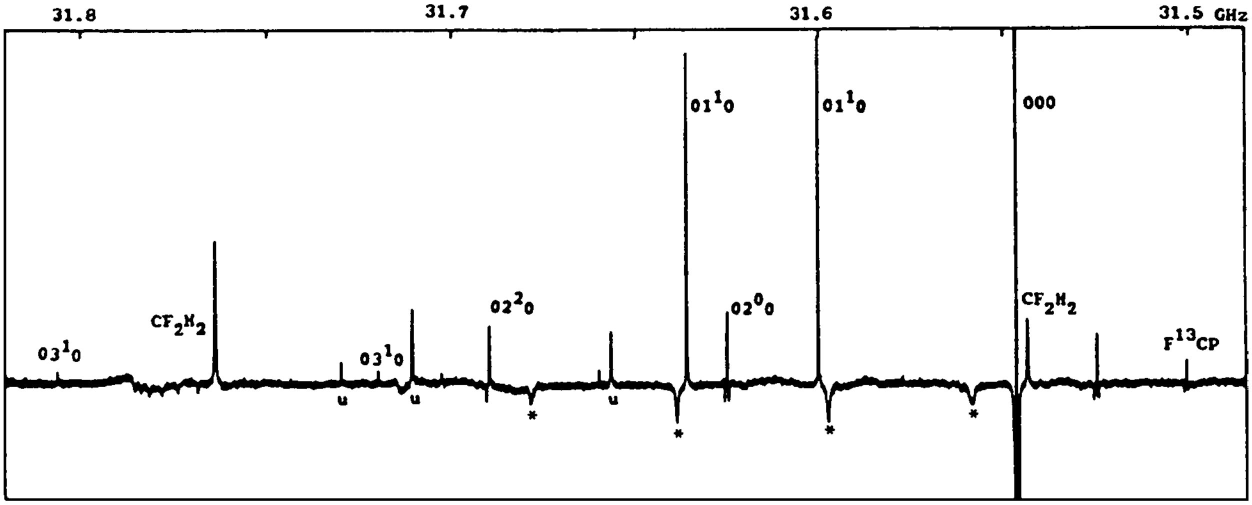


FIGURE 18 Moderate-resolution scan of the $J = 2 \rightarrow 3$ transition of FC≡P observed with a 2800-V cm^{-1} Stark modulation. Unassigned lines are labeled μ and the l-doublet Stark lobes by an asterisk. Excited-state lines from the bending mode and the splitting of the $l = \pm 1$ lines are shown. For the first excited bending state $v_2 = 1$, l-type doubling gives two $(0, 1^1, 0)$ lines shifted to high frequency relative to the ground-state line (0, 0, 0). Similar, but much weaker, lines are found for $v_2 = 3$. For the second excited state $v_2 = 2$, the lines $(0, 2^0, 0)$ and $(0, 2^2, 0)$ occur. [From Krato, H. W., Nixon, J. F., and Simmons, N. P. C. (1980). *J. Mol. Spectrosc.* **82,** 185.]

spectral lines into doublets for asymmetric rotors. With a very low barrier, such as found in CH_3NO_2, a quite complex spectrum can be obtained. In cases where asymmetric groups are connected by a single bond, internal rotation can give rise to distinct rotational isomers. These rotational isomers are often stable enough to give separate rotational spectra even though they cannot be chemically separated. As illustrated in Fig. 10, the different rotational isomers can be readily distinguished because of the very high sensitivity of the moments of inertia to the molecular geometry.

A. Single Top with a Threefold Barrier

For CH_3—CH_3 or CH_3—CHO, there are three equivalent configurations for a complete rotation of the methyl group about the C—C bond. The potential function possesses three potential energy minima and maxima as illustrated in Fig. 19. Since the methyl group is symmetric, the moments of inertia of the molecule do not depend on the orientation of the methyl group. The effects of internal rotation are transmitted to the rotational spectrum somewhat indirectly.

The periodic potential function that describes the internal rotation is expressed by

$$V(\alpha) = (V_3/2)(1 - \cos 3\alpha), \tag{79}$$

where V_3 is the threefold barrier hindering internal rotation and α the angle of internal rotation. The eigenvalue equation for this internal motion is given by

$$-F\frac{\delta^2 U(\alpha)}{\delta\alpha^2} + [V(\alpha) - E]U(\alpha) = 0, \tag{80}$$

where $V(\alpha)$ is given by Eq. (79), and $F = \hbar^2/2I_r$ with I_r the reduced moment of inertia for the relative rotation of the two groups. If the barrier is very low, $V_3 \to 0$, the form of the above waveequation is that of a spatial rotor with a fixed axis of rotation, and solution gives for the energy levels

$$E = Fm^2, \tag{81}$$

with $m = 0, \pm 1, \pm 2, \ldots$. The internal motion is essentially free rotation about the C—C bond, and the internal rotation states are specified by the quantum number m. This is the case for molecules such as CH_3NO_2, where the barrier height is 6.03 cal/mole and the rotation is effectively free (note that at room temperature, $RT \cong 600$ cal/mole). Actually, for this and other molecules of similar symmetry, there are six equivalent configurations for a complete internal revolution. The internal potential thus has sixfold symmetry, and the leading term in $V(\alpha)$ is a V_6 term:

$$V(\alpha) = (V_6/2)(1 - \cos 6\alpha), \tag{82}$$

where we expect $V_6 \ll V_3$. For the low barrier (V_6), the $m = \pm 3$ levels are particularly sensitive to the barrier.

If the barrier is very high, $V_3 \to \infty$, the internal motion of the methyl group corresponds to simple harmonic torsional oscillation in each well. The cosine function in Eq. (79) may be expanded, giving $V(\alpha) = (9V_3/4)\alpha^2$, and the form of Eq. (80) is like that for a simple harmonic oscillator. Solution gives for the energy

$$E = 3(V_3 F)^{1/2}\left(v + \tfrac{1}{2}\right), \tag{83}$$

with $v = 0, 1, 2, \ldots$. The frequency of torsional oscillation is

$$\nu = \frac{3}{2\pi}\left(\frac{V_3}{2I_r}\right)^{1/2}. \tag{84}$$

For high barriers, the rotational spectrum exhibits transitions in the ground and excited torsional states. Relative intensity measurements can thus enable the determination of the barrier. In particular, the intensity ratio between the ground and first excited states is given by the Boltzmann distribution law,

$$l_{v=0}/l_{v=1} = e^{-h\nu/kT}. \tag{85}$$

The above provides a measure of ν, and V_3 may be evaluated from Eq. (84).

For an infinite barrier, each torsional state v is threefold degenerate corresponding to oscillations in any one of the three equivalent potential wells. When the barrier is finite, the molecule may pass from one configuration to another by tunneling through the barrier. This quantum mechanical tunneling effect leads to a splitting of the threefold degeneracy into a nondegenerate level (designated by A) and a doubly degenerate level (designated by E). This torsional level splitting is illustrated in Fig. 19. Note that

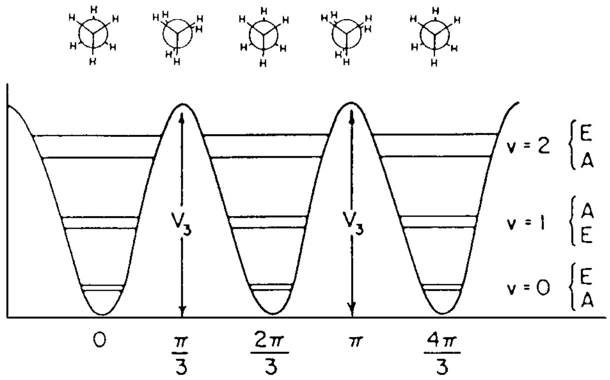


FIGURE 19 Schematic representation of the potential function and torsional energy levels for a threefold barrier. A simple cosine potential is depicted with three identical minima and maxima. The corresponding eclipsed and staggered configurations for ethane are also indicated. Each torsional energy level is labeled by the torsional quantum number v. The torsional sublevels are denoted by A or E.

the sublevel spacing increases as the torsional energy increases, and when the torsional energy is greater than the barrier height, the energy levels go over to those of a free rotor.

For intermediate values of the barrier, the torsional wavefunctions and energy levels are given by solution of the general wave equation, Eq. (80). The effects of tunneling are transmitted to the rotational spectrum by the interaction between overall and internal rotation. Since the coupling effect differs for the two torsional sublevels, there is a set of rotational energy levels associated with each sublevel, and the two sets of rotational energy levels are characterized by the Hamiltonians $\mathcal{H}_{vA}$ and $\mathcal{H}_{vE}$. For moderately high barriers, the forms of these Hamiltonians are the same as for a rigid rotor, Eq. (34), and there are two sets of rotational constants A_{vA}, A_{vE}, and so on. The effective rotational constants contain contributions from the effects of internal rotation, and A_{vA}, A_{vE}, and so on differ from each other. Hence, there is a rigid-rotorlike spectrum associated with each sublevel vA and vE. Each rotational transition in a torsional state v appears as a doublet rather than a single line. This doublet separation is a sensitive function of the barrier height, and analysis of these splittings yields the barrier that is hindering internal rotation. If the splitting cannot be observed in the ground torsional state, it can often be observed in the excited torsional states, where the splitting increases. Numerous molecules with threefold barriers have been studied by microwave spectroscopy; examples are collected in Table XVI. In addition, molecules with two equivalent methyl groups such as CH_3OCH_3 and molecules with two nonequivalent tops such as CH_3OSiH_3 have been analyzed by microwave spectroscopy. A number of molecules with twofold barriers have also been studied. Nitrobenzene is an example of this latter case.

Internal rotation analysis can be quite complicated for molecules with asymmetric rotors and asymmetric frames such as CH_2DOH. Various theoretical approaches have been developed to handle most situations. Table XVII gives a summary of barriers for various isotopic forms of CH_3OH. Depending on the molecular symmetry, some isotopic forms require V_1, V_2, and V_3 terms in the potential function. As apparent from the table, the V_3 term for —OH and —OD species is independent of isotopic substitution of the methyl group, although V_3 changes going from —OH to —OD. The changes observed provide data to understand more clearly electrostatic and other interactions characterizing hinder rotation and data to test quantum models and *ab initio* calculations.

In the application of the various theoretical approaches, the emphasis in many cases is on the derivation of an accurate potential function with as small a number of terms as possible in the Hamiltonian. On the other hand, if the desire is to provide a spectral analysis to microwave accuracy, a rather extensive number of fitting parameters is required in the Hamiltonian. For the classic internal rotation molecule CH_3OH, such an analysis provides the capabilities to calculate accurate spectral predictions over a significant range of quantum states. Such capabilities are particularly important for applications in radio astronomy. For CH_3OH, direct diagonalization of an internal rotation Hamiltonian with 60 fitting parameters, 37 of these representing interaction terms describing the coupling between torsion and end-over-end angular momentum, gave a fit of 470 rotational transitions with a root-mean-square deviation of only 0.06 MHz.

TABLE XVI Some Potential Barriers Measured with Microwave Spectroscopy

Molecule	V (cal/mole)	Molecule	V (cal/mole)
CH_3CH_2F	3306	$(CH_3)_2CO$	778
CH_3CHO	1168	$(CH_3)_2C{=}CH_2$	2210
CH_3COF	1041	$(CH_3)_3CH$	3900
CH_3COOH	483	CH_3NO_2	6.03
CH_3CH_2COOH	2360	$C_6H_5CH_3$	13.94
$CH_3CF{=}CH_2$	2440	$FC_6H_5CH_3$	13.82

TABLE XVII Barriers for Various Isotopic Forms of CH_3OH[a]

Species	V_1	V_2	V_3
CH_3OH	0	0	373.08
CD_3OH	0	0	371.8
CH_2DOH	8.80	2.42	373.22
CHD_2OH	−9.21	−2.29	373.28
CH_3OD	0	0	366.25
CD_3OD	0	0	362.19
CHD_2OD	−8.73	−2.90	365.00

[a] For details see Su, C. F., Liu, M., and Quade, C. R. (1991). *J. Mol. Spectrosc.* **146,** 264.

B. Rotational Isomerism

When the groups that rotate relative to each other are asymmetric, various rotational isomers are possible. Each isomer corresponds to a potential energy minimum. Since the moments of inertia depend on the internal rotation angle, distinct rotational spectra are observed for different isomers. For each isomer, various kinds of information including detailed molecular structures, dipole moments, and so forth may be evaluated. Furthermore, by comparing the relative intensities of the rotational lines associated with different isomers, one can obtain information on the energy difference between the isomers and determine the

more stable form. The ground-state lines of each isomer are usually accompanied by satellite lines that are from excited torsional oscillations about each potential minimum. From intensity measurements on these excited-state lines, the torsional energy level separation of the low-lying levels in the various potential wells can be evaluated. This and the above stability information can be used to specify the potential function. Such potential functions contain more terms than required for rotation of a methyl group. For *n*-propyl fluoride, $CH_3CH_2—CH_2F$, the rotation of the CH_2F group is characterized by the potential function

$$V(\theta) = (1/2) \sum_{n=1} V_n (1 - \cos n\theta), \qquad (86)$$

and the first seven potential coefficients V_n have been evaluated. The potential function for the C—C torsion is shown in Fig. 20. Two isomers are found with the *gauche* form more stable than the *trans* form. Many studies of rotational isomerism have been carried out with microwave spectroscopy; the stable forms have been identified, and an approximate potential function has been evaluated in many cases.

The power of pulsed MWFT spectroscopy in dealing with a molecule with a number of stable conformers may be illustrated by the analysis of *n*-butyl cyanide, which has three conformational isomers, as depicted in Fig. 21. Attempts to assign the spectral details by conventional techniques were unsuccessful because of the mixture of conformers and the presence of observable vibrationally excited species, which lead to a dense spectrum. On the other hand, by use of the MWFT technique, where a small amount (~1%) of *n*-butyl cyanide in Ne is expanded into the spectrometer cavity, a very low rotational temperature is produced and the molecules are forced into the ground vibrational state. The spectrum is therefore considerably simplified, and the three conformers could be assigned and analyzed. The *anti–anti* and *gauche–anti* were found to be the most stable forms.

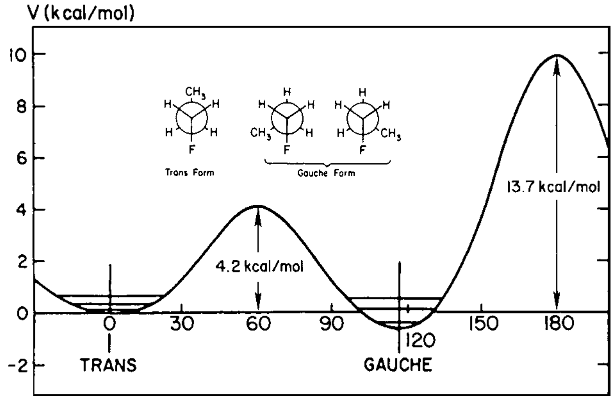


FIGURE 20 Potential function for rotation of the CH_2F group in *n*-propyl fluoride. This potential function indicates the barrier at the *cis* ($\theta = 180°$) position is higher than the barrier separating the *trans* form from the two equivalent *gauche* forms. The inset shows the conformations of *n*-propyl fluoride.

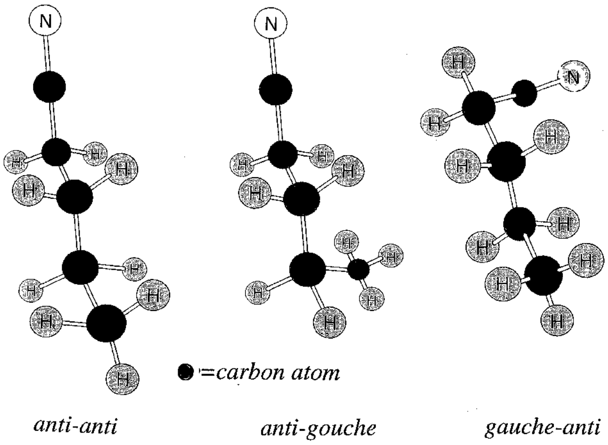


FIGURE 21 Conformational isomers of *n*-butyl cyanide. For details of the microwave analysis see Bohn, R. K., Pardus, J. L., August, J., Brupbacher, T., and Jager, W. (1997). *J. Mol. Struct.* **413,** 293.

C. Large-Amplitude Motions

When the internal motion is not well represented as a small-amplitude motion, the analysis becomes more difficult and perturbation treatments are unsatisfactory. Over the last decade, considerable improvements and advances have been made to treat the complicated spectra of certain classes of molecules. These formulations provide more convenient and accurate treatments of large-amplitude motions for (i) molecules with internal rotation, ring puckering, inversion, or umbrella-like motions, (ii) quasilinear or quasisymmetric tops, (iii) floppy molecular complexes, and (iv) molecules with two internal rotators. Description of these treatments is beyond the scope of this presentation. For further information the reader is directed to the Bibliography.

VIII. EVALUATION OF MOLECULAR STRUCTURES

A molecule is vibrating even in the ground vibrational state, and this so-called zero-point vibrational energy introduces ambiguities into structure calculations. The equilibrium structure has a well-defined physical meaning, being the distance or angle between atoms corresponding to the minimum in the potential energy. To evaluate this structure for the vibrationless state, the zero-point vibrational effects must be removed from the effective ground-state inertial constants. As noted previously, this is, in general, not possible except for relatively simple

molecules. Since the vibrational contributions are slightly different for different isotopic species, the structural parameters obtained from different combinations of isotopic species are slightly different. The spread in these effective parameters is much larger than expected from the experimental uncertainty in the moments of inertia. On the other hand, equilibrium parameters obtained from different sets of moments of inertia are consistent with the experimental uncertainties in the moments of inertia. Because of this general problem, different procedures have been developed to minimize these troublesome vibrational effects. This has led to different definitions of structural parameters depending on their method of calculation. An understanding of these different structural measures is important if comparisons are to be made of structural parameters. The different structural parameters are defined as follows:

1. Equilibrium structure r_e: the bond length or angle for the vibrationless state, evaluated by correction for the effects of vibration
2. Average structure $\langle r \rangle$ or r_z: the bond length or angle association with the average configuration of the atoms in the ground vibrational state, evaluated by partial correction for the effects of vibration
3. Effective structure r_0: the bond length or angle obtained from effective ground-state moments of inertia
4. Substitution structure r_s: the bond length or angle derived from isotopic data using differences in ground-state moments of inertia
5. Mass-dependence structure r_m: the bond length or angle derived from a large number of isotopic species by a first-order treatment of isotopic effects
6. Scaled structure r_ρ: The bond length or angle derived from scaled effective moments of inertia employing a scaling factor obtained from a set of isotopic species.

The r_0 and r_s structures are defined operationally and do not provide well-defined physical parameters. At this time the r_ρ structures provide the best near-equilibrium structures. A comparison of these different structures for SO_2 is given in Table XVIII. The evaluation of these different structures is reviewed briefly using diatomic and linear molecules as the principal examples. The values of r_e, $\langle r \rangle$, and r_0 are compared for some diatomic molecules in Table XIX.

Some of the unique structural information which can be obtained from microwave spectroscopy has been discussed in Section IV.E.

TABLE XVIII Equilibrium Structure and Various Ground-State Structures of SO_2^a

Structure	r (Å)	θ
Equilibrium (r_e)	1.4308	119°19′
Average ($\langle r \rangle$)	1.4349	119°21′
Effective (r_0)	1.4336	119°25′
Substitution (r_s)	1.4312	119°30′
Mass dependence (r_m)	1.4307	119°20′
Scaled (r_ρ)	1.4308	119°19′

A. Equilibrium Structure

By correcting B_0 for the effects of vibration, as discussed previously, one can obtain B_e. From Table II the equilibrium bond distance for a diatomic molecule X—Y is evaluated from

$$r_e = \left[\frac{m_X + m_Y}{m_X m_Y} I_b^e\right]^{1/2}, \tag{87}$$

where

$$I_b^e = \frac{h/8\pi^2}{B_e} \tag{88}$$

and the conversion factor is given by

$$h/8\pi^2 = 505{,}376 \text{ amu Å}^2 \text{ MHz}. \tag{89}$$

With more complicated molecules, additional isotopic data are needed. For a linear molecule such as XYZ, the moments of inertia for two molecular species and the expression from Table II give two equations to be solved for the two bond lengths. Alternately, the coordinates of, for example, the X-atom z_X, that is, the distance from the center of mass, can be evaluated from Kraitchman's equation (see Section VIII.D) using isotopic data from X*YZ. Subsequently, this coordinate can be used in the moment-of-inertia and first-moment equations for the XYZ species,

$$I_X = \sum m_i z_i^2, \qquad \sum m_i z_i = 0, \tag{90}$$

TABLE XIX Comparison of Various Bond Lengths (Å) for Some Diatomic Molecules

Molecule	r_e	r_0	$\langle r \rangle$
HF	0.9170	0.9257	0.9326
DF	0.9171	0.9234	0.9284
TF	0.9177	0.9230	0.9272
$H^{35}Cl$	1.2745	1.2837	1.2904
$H^{37}Cl$	1.2746	1.2837	1.2904
$D^{35}Cl$	1.2744	1.2813	1.2858
$D^{37}Cl$	1.2744	1.2813	1.2858
$T^{35}Cl$	1.2746	1.2800	1.2853
$T^{37}Cl$	1.2746	1.2800	1.2853
$^{127}I^{35}Cl$	2.3209	2.3236	2.3246
$^{127}I^{37}Cl$	2.3209	2.3235	2.3245

to give two equations in the two unknowns z_Y, z_Z. From the coordinates the bond lengths are then given by

$$d_{ij} = |z_i - z_j|. \quad (91)$$

With equilibrium moments of inertia, the important point is that the particular method of calculation is not important.

B. Average Structure

Like the equilibrium structure, the average structure has a well-defined physical meaning. The vibrational effects contained in the moments of inertia may be divided into harmonic ε_s^α(har) and anharmonic ε_s^α(anhar) contributions, which depend, respectively, on the quadratic and cubic part of the potential energy function. To evaluate the average structures, the moments of inertia for the average configuration, denoted by I_α^* ($\alpha = a, b, c$), are required. These may be obtained from the effective moments of inertia by correcting for the ε_s^α(har) effects:

$$I_\alpha^* = I_\alpha^v - \sum_s (v_s + d_s/2)\varepsilon_s^\alpha(\text{har}). \quad (92)$$

Only a knowledge of the harmonic force constants is required to make this correction. For a diatomic molecule, $\varepsilon^b(\text{har}) = -6h/8\pi^2\omega_e$; and for the ground state, $I_b^* = I_b^0 - \varepsilon^b(\text{har})/2$. The average bond length is then calculated from

$$\langle r \rangle = \left[\frac{m_X + m_Y}{m_X m_Y} I_b^*\right]^{1/2}. \quad (93)$$

In Table XIX, it is clear that $\langle r \rangle$ parameters change with isotopic substitution, but r_e parameters do not, as expected. The anharmonic part of the potential function has the effect of displacing the average configuration of a molecule from its equilibrium configuration, and $\langle r \rangle > r_e$. Usually $\langle r \rangle > r_0 > r_e$, and replacement of H by D, which significantly decreases the amplitude of vibration, causes a large shortening in $\langle r \rangle$. Average structures for excited vibrational states have also been evaluated; these clearly indicate the variation expected for a given vibrational excitation. This measure of the molecular structure is most meaningful for simple molecules. Two drawbacks are that a knowledge of the harmonic force constants is required, and if isotopic data are needed to evaluate the average structure, then the isotopic shrinkage effects just mentioned must be ignored or estimated. This introduces some ambiguity in the derived parameters.

C. Effective Structure

The effective bond distance for a diatomic molecule is given by

$$r_0 = \left[\frac{m_X + m_Y}{m_X m_Y} I_b^0\right]^{1/2}, \quad (94)$$

TABLE XX Effective and Substitution Structures of OCS

Effective structures		
	Bond length (Å)	
Isotopic species used	C—O	C—S
$^{16}O^{12}C^{32}S$, $^{16}O^{12}C^{34}S$	1.1647	1.5576
$^{16}O^{12}C^{32}S$, $^{16}O^{13}C^{32}S$	1.1629	1.5591
$^{16}O^{12}C^{34}S$, $^{16}O^{13}C^{34}S$	1.1625	1.5594
$^{16}O^{12}C^{32}S$, $^{18}O^{12}C^{32}S$	1.1552	1.5653
Average	1.1613	1.5604
Range	0.0095	0.0077

Substitution structures		
	Bond length (Å)	
Parent molecule	C—O	C—S
$^{16}O^{12}C^{32}S$	1.16012	1.56020
$^{18}O^{12}C^{32}S$	1.15979	1.56063
$^{16}O^{13}C^{32}S$	1.16017	1.56008
$^{16}O^{12}C^{34}S$	1.16075	1.55963
Average	1.16021	1.56014
Range	0.00096	0.00100

with I_b^0 the effective ground-state moment of inertia. As apparent from Table XIX, the effective bond distance r_0 varies with isotopic substitution, and $r_0 > r_e$. In general, structural parameters obtained by fitting effective moments of inertia are termed r_0 structures. For a linear XYZ molecule, two isotopic forms are required to determine the structure and also the assumption that the bond distances are unaffected by isotopic substitution. As we have seen, this assumption is only approximately true; hence, the structural parameters derived are less reliable particularly for parameters involving H atoms. When more than a minimum number of isotopic species is available, different r_0 structures can be evaluated, and an estimate of the uncertainty in the structure can be obtained (see Table XX). This analysis, however, cannot in general indicate the closeness of the r_0 to the r_e parameters.

1. Inertial Defect

The effects of vibration are readily apparent from a quantity called the inertial defect Δ, which is useful to characterize a planar molecule. The inertial defect is defined by

$$I_c^v - I_a^v - I_b^v = \Delta, \quad (95)$$

where c is the principal axis perpendicular to the molecular plane. Actually, for a planar molecule, it follows from the definitions of the moments of inertia that Δ should be identically zero. In reality this is true only if equilibrium moments of inertia are employed in the above relations.

TABLE XXI Inertial Defects for Some Planar Molecules[a]

Molecule	(amu Δ Å^2)	Molecule	(amu Δ Å^2)
H_2O	0.0486	H_2S	0.0660
H_2Se	0.0595	SO_2	0.1348
H_2CO	0.0574	$ClNO_2$	0.2079
Furan	0.046	Fluorobenzene	0.033
Pyrrole	0.076	Benzonitrile	0.084

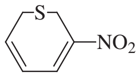
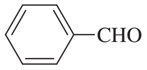
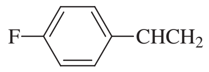
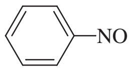
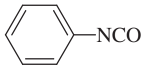
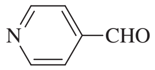
[a] $\Delta = I_c - I_a - I_b$.

When effective moments are used, the residual defined above is small but does not vanish. This arises because the vibrational effects associated with the different principal moments are slightly different. Nonetheless, a small inertial defect is usually good evidence of the planarity of a particular molecule. This is illustrated in Table XXI.

As observed, a small, positive inertial defect provides evidence for molecular planarity. On the other hand, the presence of a very low out-of-plane vibration can result in a negative inertial defects for planar molecules as illustrated in Table XXII. A simple relation has been developed to explain negative inertial defects observed for planar molecules in the ground vibrational state,

$$\Delta = -(33.715/\nu_t)\,(\text{amu Å}^2\ \text{cm}^{-1}) + \alpha\sqrt{I_c},$$

TABLE XXII Inertial Defects for Some Planar Molecules with Low-Frequency Out-of-Plane Vibrations (ν_t)[a]

Molecule	Δ (amu Å^2)	ν_t cm^{-1}
CHO—CHO	−0.1286	108
CHO—O—CHO	−0.1909	85
	−0.164	90
	−0.128	111
	−0.775	30
CH_2CH—NO_2	−0.0665	100
CHO—CHS	−0.070	104[b]
	−0.146	100
	−0.127	57
	−0.131	124

[a] $\Delta = I_c - I_a - I_b$. For an extended compilation see Oka, T. (1995). *J. Mol. Struct.* **352,** 225.
[b] Estimated.

where ν_t is the low-frequency out-of-plane vibration and I_c is the out-of-plane principal moment of inertia. The first term follows directly from theory neglecting other molecular vibrations, but it overestimates the magnitude of Δ, indicating the positive contributions from other vibrations are not negligible. Analysis of a number of molecules with low-frequency out-of-plane vibrations yields the above empirical formula with $\alpha = 0.0186$ and 0.00803 for aliphate and aromatic molecules, respectively. This expression closely approximates observed inertial defects for molecules with a low-frequency out-of-plane vibration and is hence useful in judging the planarity of molecules with small, negative inertial defects. It also follows from the above relation that ν_t can be estimated from an observed inertial object.

D. Substitution Structure

Substitution structures involve the use of Kraitchman's equations, which provide the position of an atom in a molecule utilizing the changes in moments of inertia from isotopic substitution. One isotopic form is selected as the parent molecule, and Kraitchman's equations give coordinates of the isotopically substituted atom in the center-of-mass principal axis system of the parent molecule. For diatomic or linear molecules, Kraitchman's equation has the form

$$|z| = \left[\frac{M + \Delta m}{M\Delta m}\Delta I_x\right]^{1/2}, \tag{96}$$

where $\Delta I_x = I'_x - I_x$ is the difference in the moment of inertia of the isotopically substituted molecule of mass $M + \Delta m$ and the parent molecule of mass M. The Δm is, hence, the mass change due to isotopic substitution. This relation can also be used to find the position of an atom located on the symmetry axis of a symmetric-top molecule. Only absolute values of the coordinates are obtained from relations such as Eq. (96). The sign of the coordinate must be assigned from other considerations such as the arrangement of the atoms and an approximate location of the center of mass. In general, to evaluate the bond distance between two atoms, the effective moments of inertia must be obtained for the parent molecule and two singly substituted species. Likewise, from the moment-of-inertia data for a parent and three singly substituted species, an interatomic angle can be evaluated. For a diatomic molecule, the parent X—Y and two isotopic forms X*—Y and X—Y* are required to give the internuclear distance

$$d_{X-Y} = |z_X - z_Y|. \tag{97}$$

Since differences in the effective moments of inertia are employed in calculating substitution structures, the

zero-point vibrational effects tend to cancel, and more consistent structural parameters can be obtained from different combinations of isotopic data. This is illustrated in Table XX by the different r_s structures for OCS. Note that the spread in the r_s structures is 10 times smaller than the spread in the r_0 structures. One of the best approximations to the equilibrium structure is the substitution structure, and numerous such structures have been determined, with typical structural uncertainties of ±0.002 to ±0.005 Å and ±0.2° to ±0.5°.

E. Mass-Dependence Structure

To evaluate the mass-dependence structure, various isotopic species are employed to calculate the substitution coordinates of each nonequivalent atom for a given parent. These substitution coordinates are then used to evaluate the moments of inertia, Eq. (5), which are called substitution moments of inertia I_α^{s}. The mass-dependence moment of inertia I_α^{m} is calculated from the relation

$$I_\alpha^{\mathrm{m}} = 2I_\alpha^{\mathrm{s}} - I_\alpha^0 \qquad (\alpha = a, b, c). \tag{98}$$

To first order, the I_α^{m} moments are equal to the equilibrium moments of inertia I_α^{e}. The above procedure is repeated for another parent species. Once a sufficient number of I^{m} for different parent isotopic species have been determined, the moment-of-inertia equations may be solved to give the r_{m} structure. The r_{m} structure for SO_2 is given in Table XVIII. This measure of the molecular structure has limited applicability because of the large amount of precise isotopic moment-of-inertia data required and because the first-order approximation $I^{\mathrm{m}} \cong I^{\mathrm{e}}$ is not sufficient especially for light atoms. Thus, hydrogen bond lengths cannot be determined by this method.

F. Scaled Structure

A method has been proposed which employs a set of moments $[I^\rho]_g$ defined for L isotopic species and computed from

$$[I^\rho]_g = (2\rho - 1)[I^0]_g, \qquad g = 1, 2, \ldots, L, \tag{99}$$

with the scaling factor obtained from

$$\rho = [I^{\mathrm{s}}]_1 / [I^0]_1. \tag{100}$$

Here $[I^0]_1$ $(g = 1)$ is the ground-state moment of the parent isotopic species and $[I^{\mathrm{s}}]_1$ is the substitution moment of inertia calculated from the set of substitution coordinates of the parent species. The dataset, the L isotopic species, is that needed for evaluation of a complete substitution structure. For a linear XYZ molecule, four isotopic species (including the parent) are needed for the substitution structure. From Eqs. (99) and (98), it follows for the parent, $g = 1$, that $[I^\rho]_1 \equiv I^m$. The rationale for this method is primarily based on the observation that ρ is constant for different choices of parent species to ca. $1/10^4$ $([I^{\mathrm{s}}]_g/[I^0]_g \cong [I^{\mathrm{s}}]_1/[I^0]_1)$, and the I^ρ computed from the above prescription give values very close to I^{e}.

Once the set of L scaled moments of inertia I^ρ have been evaluated, the molecular structural parameters are derived by means of a standard least-squares fitting of the I^ρ's. This is found to provide the best averaging of small residual vibrational effects. For a linear triatomic molecule XYZ, the four moments of inertia would be analyzed for the parameters d_{YX} and d_{YZ} (see Table II). Importantly, the method employs a minimal set of isotopic substitution data compared to the mass-dependence method. It is, however, necessary to select the parent such that all isotopic substitutions satisfy either $\Delta m_i > 0$ or $\Delta m_i < 0$ for all atoms i. This minimizes residual vibrational effects. For the general case, there are moments I_α^0 and I_α^{s} $(\alpha = a, b, c)$ associated with each axis, and these are used to calculate the corresponding ρ_a, ρ_b, ρ_c and the I_a^ρ, I_b^ρ, I_c^ρ. The moments of inertia I_α^ρ are then analyzed by least squares for the structural parameters. Table XXIII compares several structures for OCS. Results for SO_2 are summarized in Table XVIII. It is apparent that the r_ρ structures compare most favorably with the r_{e} structures. Similar results are found for other molecules.

This method based on the use of a set of scaled mometns of inertia provides a molecular structure which is a better approximation to the r_{e} structure than the conventional r_{s} structure, particularly for heavy-atom molecules. For molecules which contain hydrogen atoms, additional considerations apply because of larger vibrational effects. The quantity ρ is no longer virtually constant, but varies significantly when hydrogen atom substitution species $(H \leftrightarrow D, T)$ are considered, and the I^ρ values do not give reliable structures. Corrections now have to be applied to obtain near-equilibrium structures, and an empirical method has been developed. In particular, the moments of inertia for the deuterated species are corrected via

$$(I^\rho)^{\mathrm{D}}_{\mathrm{corr}} = (I^\rho)^{\mathrm{D}} + \Delta. \tag{101}$$

TABLE XXIII Structural Calculations for Carbonyl Sulfide (OCS)[a]

	r_s	r_0	r_m	r_ρ	r_e
CO	1.1605	1.1568	1.1587	1.1551	1.1543
CS	1.5596	1.5645	1.5593	1.5621	1.5628

[a] All distances in angstroms. [From Harmony, M. D., and Taylor, W. H. (1988). *J. Mol. Spectrosc.* **118,** 163; see also Berry, R. J., and Harmony, M. D. (1988). *Struct. Chem.* **1,** 49.]

The correction factor Δ utilizes a bond elongation parameter δr_D ($\cong$0.003 Å). Simple correction procedures have been given for various molecule types. The $(I^\rho)^D_{corr}$ and the other scaled moments of the complete set are then used in a least-squares determination of the structural parameters. The results found are quite good. The scaled method represents a significant step in the improvement of spectroscopic structure determinations of polyatomic molecules.

IX. STARK EFFECT AND ZEEMAN EFFECT

Applied electric or magnetic fields modify the rotational spectrum. The most important of these is the effect of electric fields, commonly called the Stark effect. In the absence of an electric field, the rotational states are $(2J+1)$-fold degenerate since the energies are independent of M,

$$M = 0, \pm 1, \pm 2, \ldots, \pm J, \qquad (102)$$

which specifies the possible orientations of the angular momentum vector relative to a fixed direction in space. When an electric field is present, the field interacts with the molecular dipole moment and the rotational energy levels split into a number of sublevels. The degeneracy in the space orientation quantum number M is thus partially or completely lifted by the interaction. A schematic illustration is provided in Fig. 3. A rotational line, therefore, splits into a number of components. The general selection rules for rotational transitions are

$$M \to M, \quad M \to M \pm 1, \qquad (103)$$

in addition to those already given for the different types of rotors (Section IV). Usually, the electric field is applied parallel to the electric vector of the radiation, and the selection rule $\Delta M = 0$ applies.

The Stark effect is usually used to modulate rotational lines to improve their detection. This is the basis of the Stark-modulation spectrometer (Fig. 2) discussed earlier. With this type of spectrometer both the Stark lines and the zero-field transitions are displayed. The Stark effect pattern can be a valuable aid in the assignment of rotational spectra, particularly for asymmetric tops. Specifically, by counting the number of Stark components, one can obtain an indication of the smaller J value involved in the transition. Another particularly important application of this effect is in the evaluation of very accurate electric dipole moments. These are determined by careful measurement of the displacement of the Stark components from the zero-field absorption line as a function of the applied field. The analysis of these Stark splittings by means of the appropriate expression allows evaluation of the dipole components along the principal axes.

In the general case, the molecular dipole moment fixed in the molecule can have three nonvanishing components μ_x, μ_y, μ_z along the molecule-fixed principal axis system. The Stark effect Hamiltonian is then expressed by

$$\mathcal{H}_{\mathcal{E}} = -\mathcal{E} \sum \mu_g \Phi_{Zg}, \qquad g = x, y, z, \qquad (104)$$

where Φ_{Zg} are the direction cosines of the x, y, z axes relative to the space-fixed Z axis, the direction along which the electric field $\mathcal{E}$ is applied. For linear and symmetric-top molecules only a dipole component along the z axis is present: $\mu_z = \mu$ and $\mu_x = \mu_y = 0$. Usually, the above interaction can be adequately treated by first- and second-order perturbation theory. When levels that interact via $\mathcal{H}_{\mathcal{E}}$ are degenerate or near degenerate, the frequency displacement $\Delta\nu$ of the Stark components is linear in the electric field and depends on M, that is, $\Delta\nu = B_{J\tau} M \mathcal{E}$. Here $B_{J\tau}$ are functions of J and other possible quantum numbers. In this case, a rotational line splits into $2J+1$ components. On the other hand, when levels interact via $\mathcal{H}_{\mathcal{E}}$ which are widely separated, the frequency displacement varies as $\Delta\nu = (A_{J\tau} + B_{J\tau} M^2)\mathcal{E}^2$, and only $J+1$ components are obtained for $\Delta J = \pm 1$ and J components for $\Delta J = 0$. Here J refers to the smaller J value involved in the transition.

The relative intensities of the Stark components depend on the transition type

$$\left.\begin{array}{ll} I_M = P[(J+1)^2 - M^2], & J \to J+1 \\ I_M = QM^2, & J \to J \\ I_M = R[J^2 - M^2], & J \to J-1 \end{array}\right\} M \to M. \qquad (105)$$

For asymmetric tops all three expressions apply, while for linear and symmetric-top molecules only the first intensity expression is needed. The P, Q, R coefficinets are independent of M but depend on the intensity of the unsplit line. Note that the $M = 0$ component is forbidden for a $\Delta J = 0$ transition, and for a second-order effect ($\Delta\nu \sim M^2$), a factor of $1/2$ must be included in the intensity expression for the $M = 0$ component since the $+M$ and $-M$ degeneracy is lost.

The general features of the second-order effect for the linear molecule OCS are illustrated in Fig. 3. Clearly, as the electric field is increased, the field-dependent Stark components move further away from the zero-field line. The three components $|M| = 0, 1, 2$ are as expected for a $J = 2 \to 3$ transition. The component nearest the unsplit line requires a high field before it shows up clearly. By plotting $\Delta\nu$ against $\mathcal{E}^2$, the effect is easily classified as a second-order effect. The appearance of a first-order Stark

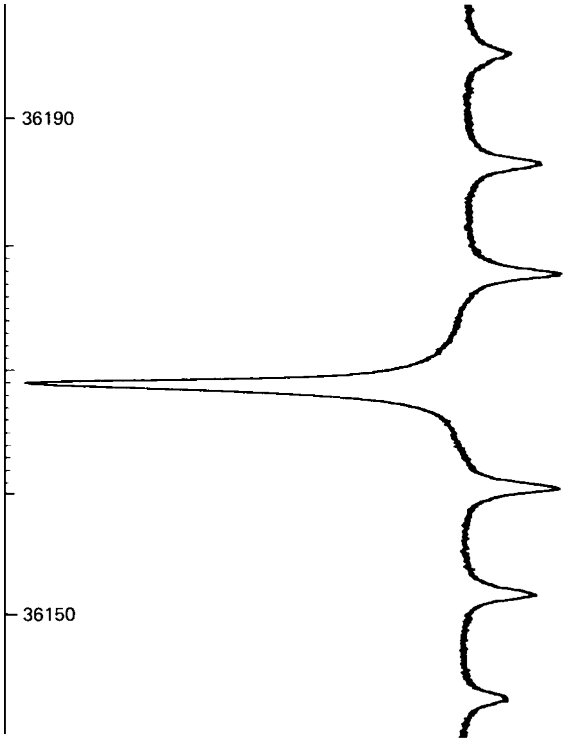


FIGURE 22 A methanol transition recorded with 175-V Stark modulation. The Stark pattern is typical of a first-order effect.

effect, where the Stark lobes form a symmetrical pattern about the zero-field line, is illustrated in Fig. 22 for a transition of CH_3OH.

In addition to the Stark effect produced by the permanent dipole moment, small electric dipole moments are induced by the electric field, and this effect is characterized by the electric-polarizability tensor α. This is important only with very large electric fields and will not be discussed further.

A. Linear and Symmetric-Top Molecules

We now consider the Stark effect quantitatively for linear and symmetric-top molecules. The first-order effect possible for a symmetric top is given by standard first-order perturbation theory as

$$\left[E^{(1)}_{\mathscr{E}}\right]_{JKM} = (J, K, M|\mathscr{H}|_{\mathscr{E}} J, K, M) = -\mu\mathscr{E}(J, K, M|\Phi_{Zz}|J, K, M) = \frac{-\mu\mathscr{E}KM}{J(J+1)} \tag{106}$$

for the level J, K, M, where the matrix element is evaluated in the symmetric rotor basis $|J, K, M)$. The rotational motion is like that of a top spinning about its symmetry axis and also precessing about the vertical Z axis. The frequency displacement is found by applying the selection rules $J \rightarrow J+1$, $K \rightarrow K$, and $M \rightarrow M$:

$$\Delta\nu^{(1)}(\Delta M = 0) = (0.50344)\mu\mathscr{E}\left[\frac{2KM}{J(J+1)(J+2)}\right], \tag{107}$$

where J is the quantum number of the lower state of the transition, and the factor 0.50344 has been introduced so that $\Delta\nu$ is in MHz units if μ is in debye units and $\mathscr{E}$ is in volts per centimeter. Table XXIV gives the values of the quantity in brackets for some low J values. For example, with $\mu = 1$ D and $\mathscr{E} = 300$ V/cm, we find for the $(J, K, M) = (1, 1, 1) \rightarrow (2, 1, 1)$ transition that $\Delta\nu^{(1)} = 50$ MHz, a rather large splitting. Since K and M in Eq. (107) can each take on values between J and $-J$, a symmetric Stark pattern is produced. There is also a much smaller second-order effect that must be considered if $\mu\mathscr{E}$ is large.

When K or M is zero, the first-order effect vanishes, and only a quadratic effect is found, as for a linear molecule. The second-order effect calculated for a linear molecule also applies to a symmetric top when K is zero. Standard second-order perturbation theory gives for the Stark correction of the J, M level,

$$\left[E^{(2)}_{\mathscr{E}}\right]_{J,M} = \sum_{J'}{}' \frac{|(J, M|\mathscr{H}_{\mathscr{E}}|J', M)|^2}{E_J - E_{J'}} = \mu^2\mathscr{E}^2\left[\frac{|(J, M|\Phi_{Zz}|J+1, M)|^2}{E_J - E_{J+1}} + \frac{|(J, M|\Phi_{Zz}|J-1, M)|^2}{E_J - E_{J-1}}\right]. \tag{108}$$

TABLE XXIV Relative First-Order Stark Shifts for Symmetric Tops[a]

	Value of $2KM/J(J+1)(J+2)$ for various (K, M)									
	(1, 1)	(2, 2)	(2, 1)	(3, 3)	(3, 2)	(3, 1)	(4, 4)	(4, 3)	(4, 2)	(4, 1)
$J = 1 \rightarrow 2$	0.3333	—	—	—	—	—	—	—	—	—
$J = 2 \rightarrow 3$	0.0833	0.3333	0.1666	—	—	—	—	—	—	—
$J = 3 \rightarrow 4$	0.0333	0.1333	0.0666	0.3000	0.2000	0.1000	—	—	—	—
$J = 4 \rightarrow 5$	0.0166	0.0666	0.0333	0.1500	0.1000	0.0500	0.2666	0.2000	0.1333	0.0666

[a] Multiply entry by $(0.50344)\mu\mathscr{E}$ to obtain the shift of the Stark component from the undisplaced line.

TABLE XXV Relative Second-Order Stark Shifts for Linear Molecules[a]

	Value of $\Delta[J(J+1)-3M^2]/J(J+1)(2J-1)(2J+3)$ for various M				
	0	1	2	3	4
$J=0\to 1$	0.5333	—	—	—	—
$J=1\to 2$	−0.1524	0.1238	—	—	—
$J=2\to 3$	−0.0254	−0.0071	0.0476	—	—
$J=3\to 4$	−0.0092	−0.0056	0.0052	0.0288	—
$J=4\to 5$	−0.0044	−0.0033	−0.0001	0.0054	0.0165

[a] Multiply entry by $(0.50344)^2\mu^2\mathscr{E}^2/2B$ to obtain the shift of the Stark component from the undisplaced line.

The perturbation connects adjacent J states. Evaluation of the direction cosine matrix elements in the basis $|J, M)$ and the energy level differences gives

$$\left[E^{(2)}_{\mathscr{E}}\right]_{J,M} = (0.50334)^2 \times \frac{\mu^2\mathscr{E}^2}{2B}\left[\frac{J(J+1)-3M^2}{J(J+1)(2J-1)(2J+3)}\right], \quad J\neq 0 \qquad (109)$$

In addition, for $J = 0$ and $M = 0$, we have

$$\left[E^{(2)}_{\mathscr{E}}\right]_{J=0} = -(0.50344)^2\mu^2\mathscr{E}^2/6B. \qquad (110)$$

The Stark splitting $\Delta\nu^{(2)}$ depends on the difference of the bracket term in Eq. (109) for the two levels involved in the transition. Table XXV gives the difference in the bracket term for some $J\to J+1$, $M\to M$ transitions. It follows from the table that Stark components can occur on both the high- and low-frequency sides of the unperturbed line. For the $J=2\to 3$ transition, two low-frequency Stark lobes and one high-frequency lobe are predicted and observed for OCS in Fig. 3. Furthermore, the magnitude of the splitting decreases with increasing J, and the largest high-frequency displacement occurs for $M=J$. As a quantitative example, consider OCS, where $2B=12{,}162.97$, $\mu=0.715$ D, and a large field $\mathscr{E}=2800$ V/cm. For the $J, M=(2, 2)\to(3, 2)$ transition, we find $\Delta\nu^{(2)}=4$ MHz, which is easily observable but quite small compared with a first-order effect.

B. Asymmetric-Top Molecules

For an asymmetric rotor, the analysis basically follows the same procedure; however, the details require more space than possible in the scope of this presentation. The Stark effect is usually second order, and the Stark energy contains contributions from all three dipole components (if present). The contribution of each component has the form

$$\left[E^{(2)}_{\mathrm{g}}\right]_{J\tau M} = \mu_{\mathrm{g}}^2\mathscr{E}^2\left[A_{J,\tau} + B_{J,\tau}M^2\right], \qquad (111)$$

where A and B are called Stark coefficients and depend on J, τ, and the inertial asymmetry κ. Though simple expressions cannot be given for these coefficients, they may be calculated from second-order perturbation theory and knowledge of the direction cosine matrix elements in the asymmetric rotor basis. They have been tabulated for low J. Once these quantities are specified, the second-order Stark energy may be calculated from Eq. (111).

With asymmetric rotors, the possibility of degeneracies or near degeneracies exists, and in this case the above second-order expression does not apply. Near degeneracies often occur between asymmetry doublets. For $J=2$ and a near-prolate rotor, for instance, the pair of levels $2_{1,2}$, $2_{1,1}$ or $2_{2,1}$, $2_{2,0}$ interact via μ_a, and this interaction is very large if the levels are very close together. Ordinary second-order perturbation theory then fails. Thus a transition involving a level that can interact with a nearby level does not exhibit a typical second-order effect. To a good approximation, these levels may be separated from the other levels, and the problem can be treated by standard methods of quantum mechanics as a two-level system. Consider two levels ψ_1^0, ψ_2^0 that are eigenfunctions of the unperturbed Hamiltonian $\mathscr{H}_0$ with eigenvalues E_1^0 and E_2^0. The complete Hamiltonian is $\mathscr{H}=\mathscr{H}_0+\mathscr{H}_{\mathscr{E}}$, with $\mathscr{H}_{\mathscr{E}}$ the perturbation. In the basis of $\mathscr{H}_0$, there are no off-diagonal elements from $\mathscr{H}_0$ and no diagonal elements for $\mathscr{H}_{\mathscr{E}}$. The secular determinant thus has the form

$$\begin{vmatrix} E_1^0 - E & \mathscr{E}\mu_{12} \\ \mathscr{E}\mu_{12} & E_2^0 - E \end{vmatrix} = 0, \qquad (112)$$

where $\mu_{12}\mathscr{E} = (1|\mathscr{H}_{\mathscr{E}}|2)$. The roots are

$$E_{1,2} = \tfrac{1}{2}\left(E_1^0 + E_2^0\right) \pm \tfrac{1}{2}\left[\left(E_1^0 + E_2^0\right)^2 + 4\mathscr{E}^2\mu_{12}^2\right]^{1/2}, \qquad (113)$$

with $E_1^0 > E_2^0$; the plus sign is used for E_1 and the negative sign for E_2. When $(E_1^0 - E_2^0)^2 > 4\mathscr{E}^2\mu_{12}^2$, the above can be expanded to give a second-order effect:

$$E_1 = E_1^0 + \frac{\mu_{12}^2\mathscr{E}^2}{E_1^0 - E_2^0} + \cdots, \qquad (114)$$

$$E_2 = E_2^0 - \frac{\mu_{12}^2\mathscr{E}^2}{E_1^0 - E_2^0} + \cdots. \qquad (115)$$

Note, however, that the second-order effect could be rather large if the energy denominator is small. If the perturbation is large, $(E_1^0 - E_2^0)^2 < 4\mu_{12}^2\mathscr{E}^2$, then

$$E_1 = \tfrac{1}{2}\left(E_1^0 + E_2^0\right) + \mu_{12}\mathcal{E}, \qquad (116)$$

$$E_2 = \tfrac{1}{2}\left(E_1^0 + E_2^0\right) - \mu_{12}\mathcal{E}, \qquad (117)$$

which is a first-order effect. Once the matrix element is specified $\mu_{12} = \mu KM/J(J+1)$, the above relations provide the Stark energies. Such effects are possible not only for asymmetric tops, they also can be observed for linear molecules in excited bending modes, where μ_{12} involves the quantum number l rather than K.

C. Dipole Moment Information

By analysis of the Stark effect, very accurate dipole moments can be evaluated, and many hundreds of molecules have now been studied. Illustrative dipole moments determined by microwave spectroscopy are collected in Table XXVI. Some of the interesting effects observed and the implications will be briefly discussed. Since the dipole moment obtained is associated with a particular isotopic species, vibrational state, and rotational transition, it is possible to observe small changes in μ with these quantities. In LiCl, the dipole moment is found to increase with vibrational state $\mu(v=0) = 7.120$ D, $\mu(v=1) = 7.20$ D, and $\mu(v=2) = 7.296$ D. This is consistent with a largely ionic bond, where the average separation between the atoms increases with v. Because deuterium substitution tends to decrease the average separation between the atoms, a decrease in the dipole moment may be expected.

TABLE XXVI Illustrative Dipole Moments

Molecule	μ (D)	Molecule	μ (D)
	Linear molecules		
LiCl	7.119	CO	0.112
HCN	2.985	HCP	0.390
OCS	0.7152	HCCF	0.73
	Symmetric-top molecules		
CH_3F	1.854	CH_3Cl	1.869
CH_3Br	1.797	CH_3l	1.618
CH_3CN	3.913	CH_3CCH	0.784
$(CH_3)_3N$	0.612	$(CH_3)_3P$	1.192

Asymmetric-top molecules

Molecule	μ_a	μ_b	μ_c	$\mu(D)$
H_2CO	—	—	—	2.339
HFCO	0.58	1.91	—	1.99
HCOOH	1.391	0.26	—	1.415
CH_3CFO	2.83	0.88	—	2.96
trans-$CH_3CH_2CH_2F$	1.970	0.566	—	2.050
gauche-$CH_3CH_2CH_2F$	1.137	1.450	0.472	1.902
CH=CHF	1.280	0.629	—	1.427
$(CH_3)_2PH$	—	1.15	0.44	1.23

This is found in a number of cases, but in HCP, for example, deuterium substitution increases the moment by 0.007 D. This reflects the fact that the electronic structure can also be affected by isotopic substitution. The use of deuterium substitution to convert a nonpolar molecule to a slightly polar molecule can be illustrated by CH_3CD_3. The isotope-induced moment is about 0.01 D, and microwave spectra for this molecule have been observed. The effects of centrifugal distortion have also been observed. The near-prolate asymmetric rotor HNCO has an exceptionally large energy of rotation about the a axis, which is essentially along the NCO chain. The NCO chain is essentially linear and the HNC angle is 128°. The moment μ_a is found to decrease with increasing K. The values are $\mu_a(K=0) = 1.577$ D, $\mu_a(K=1) = 1.572$ D, $\mu_a(K=2) = 1.540$ D, and $\mu_a(K=3) = 1.500$ D for the $J = 3 \rightarrow 4$ transition. This may be explained as follows: as K increases, the rotation about the a axis increases and the HNC angle decreases due to the centrifugal force; thus the projection of the N—H bond moment on the a axis decreases and hence μ_a. As a final example, we mention the use of microwave Stark effect measurements to determine the dipole moment direction in the molecule. The Stark effect does not provide the signs of the dipole components, and therefore the direction must be obtained from other information, such as electronegativities. However, the effect of isotopic substitution, where the primary effect is to rotate the principal axis system, has been used to specify the directions of the dipole components and hence μ. This method is illustrated in Fig. 23.

D. Zeeman Effect

Applied magnetic fields affect the microwave spectrum in a manner analogous to the Stark effect. This is called

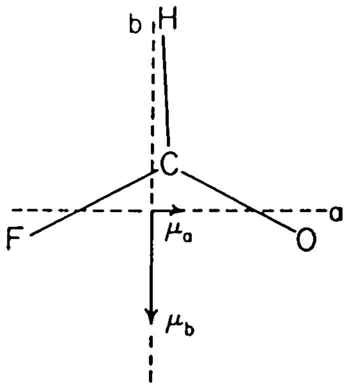


FIGURE 23 Principal inertial axes of HCOF. The direction of μ_b may be selected from chemical arguments, but not that of μ_a. However, for μ_a to increase on deuteration, which results in a rotation of the b axis toward the D atom, the direction must be as shown. The arrowhead indicates the negative end of the dipole moment components.

the Zeeman effect. Except for molecules with nonzero electronic angular momentum and consequently a permanent magnetic dipole moment, the effect is small, but observable nonetheless with kilogauss magnetic fields. The Zeeman effect Hamiltonian for the interaction of a magnetic dipole with an applied field H is given by

$$\mathcal{H}_H = -\boldsymbol{\mu} \cdot \mathbf{H}. \tag{118}$$

For molecules in singlet ground electronic states, the magnetic moment μ is generated by the molecular rotation (rotation of charges). With the assumption that the molecular magnetic moment is proportional to the angular momentum, it can be shown that the Zeeman splitting of the rotational levels is given by

$$E_H = -g_{J,\tau}\beta_I HM, \tag{119}$$

where $g_{J,\tau}$ is the rotational g-factor and depends on the inner quantum numbers K and τ for symmetric and asymmetric tops, respectively. Here β_I is the nuclear magneton, H is the applied magnetic field, and M is the orientation quantum number. The selection rules are $\Delta M = 0$ or ± 1 depending on whether the magnetic field is applied parallel or perpendicular to the electric radiation vector. Analysis of the Zeeman splittings provides values of the molecular rotational g-factors. These g-factors, for example, when measured for two isotopic species, can be used to determine the vector direction of the electric dipole moment.

Besides the molecular magnetic moments generated by rotation, smaller magnetic moments are induced by the external field. These additional effects are expressed in terms of a magnetic susceptibility tensor $\boldsymbol{\chi}$, and analysis of these effects yields the elements of the magnetic susceptibility tensor. Considerable progress has been made in the study of magnetic properties of molecules and in deriving information on the electronic structure of molecules from these studies. However, their description is outside the scope of this presentation.

X. NUCLEAR QUADRUPOLE HYPERFINE STRUCTURE

The effects of nuclear coupling can give rise to hyperfine structure in the rotational spectrum. The most important type of nuclear interaction occurs when a nucleus with a nonzero nuclear quadrupole moment ($I > 1/2$) is present in the molecule. The nuclear quadrupole moment Q measures the deviation of the nuclear charge distribution from a spherical distribution. Generally, Q becomes larger as the nucleus becomes heavier. A positive Q indicates a prolatelike distribution (elongated along the spin axis) and a negative Q indicates an oblatelike distribution (flattened along the spin axis). This nonspherical nuclear charge can interact with the surrounding unsymmetrical molecular charge distribution arising from all the other charges in the molecule. This latter distribution is measured by the electric field gradient $\partial^2 V/\partial^2 z^2$, with V the electrostatic potential. In such cases the nuclear spin $\mathbf{I}$ is coupled to the overall rotation $\mathbf{J}$, and a nuclear quadrupole hyperfine structure results. The total angular moment $\mathbf{F} = \mathbf{J} + \mathbf{I}$ has the magnitude $\hbar\sqrt{F(F+1)}$, and a new quantum number F is now required to characterize the energy levels, or hyperfine states,

$$F = J + I, J + I - 1, J + I - 2, \ldots, |J - I|. \tag{120}$$

Each rotational level is hence split into a number of different levels labeled by the values of F. When $J > I$, there are in general $(2I + 1)$ values of F. The projection of F along an axis fixed in space, $\hbar M_F$, is specified by the quantum number

$$M_F = F, F - 1, F - 2, \ldots, -F. \tag{121}$$

This quantum number becomes important when an external field is applied to a molecule with a quadrupolar nucleus. Since the rotational states are split by the nuclear interaction, a given rotational transition splits into a number of components and a group of closely spaced lines is observed. The selection rules are as before, with the additional rules

$$F \to F, \qquad F \to F \pm 1, \qquad I \to I. \tag{122}$$

This splitting, in general, increases for some of the more common nuclei in the order ^{14}N$(I = 1) <$ ^{37}Cl$(I = 3/2) <$ ^{35}Cl$(I = 3/2) <$ ^{81}Br$(I = 3/2) <$ ^{79}Br$(I = 3/2) <$ ^{127}I$(I = 5/2)$. Many common nuclei have $I = 0$ (e.g., ^{12}C, ^{16}O, ^{32}S) or $I = 1/2$ (e.g., ^{1}H, ^{13}C, ^{15}N, ^{19}F) and hence do not give rise to a quadrupole coupling interaction. The quadrupole splittings decrease with increasing J, often becoming unresolvable at sufficiently high J. Nuclear quadrupole hyperfine structure in the rotational spectrum indicates the presence of a quadrupolar nucleus such as Cl; it can be useful in the assignment of a spectrum and it provides information on chemical bonding.

A. Linear and Symmetric-Top Molecules

The nuclear quadrupole energy for a linear molecule with a single coupling nucleus in the absence of external fields is given by

$$E_Q = -\chi Y(J, I, F), \tag{123}$$

where $\chi = eQq$ is the nuclear quadrupole coupling constant in frequency units with $q = \partial^2 V/\partial z^2$, the electric

field gradient relative to the principal axis system, and where z is along the molecular axis and Q is considered known and is characteristic of the coupling nucleus. Also,

$$Y(J, I, F) = \frac{(3/4)C(C+1) - I(I+1)J(J+1)}{2(2J-1)(2J+3)I(2I-1)} \tag{124}$$

$$C = F(F+1) - J(J+1) - I(I+1) \tag{125}$$

is a function of the various quantum numbers and the spin I. It has been tabulated for various I and J. It may be observed that $Y(J, I, F)$ is undefined for $I = 0$ and 1/2, which is consistent with the requirement $I \geqq t$ for a nuclear quadrupole interaction. The E_Q must be added to the rigid-rotor energy E_r to give the total energy. Applying the selection rules $\Delta J = 0$, $\Delta F = 0, \pm 1$, one obtains the rotational frequencies including effects of quadrupole coupling:

$$\nu = \nu_r - \chi[Y(J+1, I, F') - Y(J, I, F)], \tag{126}$$

where ν_r is the unperturbed rotational frequency and $F' = F$, $F \pm 1$. The coupling constant can be evaluated from the splitting between any two hyperfine components. To evaluate the rotational constant, the rigid-rotor frequency ν_r is required, and this may be evaluated by correcting the hyperfine components ν with the known χ via the above frequency expression.

To understand the appearance of the hyperfine pattern, a knowledge of the relative intensities of the components is required. The explicit expressions require too much space to give here. However, we may point out that for any class of rotor when $J > I$, there are $2I + 1$ components for $F \to F + 1$, $2I$ components for $F \to F$, and $2I - 1$ components for $F \to F - 1$. Furthermore, the most intense components are those where $\Delta F = \Delta J$. An approximate intensity rule is that for $J \to J + 1$ transitions, the intensities of the $F \to F + 1$ components are proportional to F, while the $F \to F$ components are considerably weaker and the $F \to F - 1$ components even weaker. For the $J \to J$ transition, the intensity of the $F \to F$ component is proportional to F, while the $F \to F \pm 1$ components are considerably weaker. In fact, the intensities of the components for $F \neq \Delta J$ decrease rapidly with increasing J. Also, the function $Y(J, I, F)$ may be positive or negative, but for the maximum and minimum values of F, the function is positive. Moreover, the strongest component is usually not significantly displaced from the rigid-rotor position. Consider, for example, the $2 \to 3$ transition with $I = 3/2$. For $J = 2$, $F = 1/2, 3/2, 5/2, 7/2$ and for $J = 3$, $F = 3/2, 5/2, 7/2, 9/2$. We expect four $F \to F + 1$ components ($1/2 \to 3/2$, $3/2 \to 5/2$, $5/2 \to 7/2$, and $7/2 \to 9/2$ in order of increasing intensity), three weaker $F \to F$ components ($3/2 \to 3/2$, $5/2 \to 5/2$, and $7/2 \to 7/2$), and two even weaker $F \to F - 1$ components ($5/2 \to 3/2$ and $7/2 \to 5/2$). These considerations enable the identification of the hyperfine components of a given rotational transition. Early applications of microwave spectroscopy to the study of hyperfine structure used the appearance of the fine structure, that is, the relative spacings and intensities of the components, to determine unknown nuclear spins. For example, the spin of ^{33}S was found to be 3/2 by this method.

For a symmetric top, with the coupling atom on the symmetry axis,

$$E_Q = \chi \left[\frac{3K^2}{J(J+1)} - 1 \right] Y(J, I, F), \tag{127}$$

where χ is the coupling constant with reference to the molecular axis of symmetry. From the above expression it follows that each J, K level splits into a number of sublevels of different F. The selection rules are $\Delta J = \pm 1$, $\Delta K = 0$, $\Delta I = 0$, $\Delta F = 0, \pm 1$. When $K = 0$, the hyperfine pattern is like that for a linear molecule. For other K values, a similar pattern is obtained. However, when different K lines are separated by less than the quadrupole splitting, the individual patterns for each K overlap, and a quite complex overall structure can be obtained.

For coupling atoms off the symmetry axis we have a more complicated problem. An example would be $HCCl_3$. The hyperfine structure for molecules with two or more coupling nuclei is more complex but has been treated theoretically and observed experimentally.

B. Asymmetric-Top Molecules

For an asymmetric prolate rotor with a single coupling nucleus we have

$$E_Q = \left\{ \left[3\langle P_a^2 \rangle - J(J+1) \right] \chi_{aa} - \sigma \left[\langle P_a^2 \rangle - W(b_p) \right] \eta \chi_{aa} \right\} \times \frac{Y(J, I, F)}{J(J+1)}, \tag{128}$$

where $\sigma = -1/b_p$, and $W(b_p)$ is Wang's reduced energy and $\langle P_a^2 \rangle$ the average P_Z^2 in the asymmetric rotor basis. The asymmetry parameter

$$\eta = \frac{\chi_{bb} - \chi_{cc}}{\chi_{aa}} \tag{129}$$

measures the departure of the field gradient from cylindrical symmetry about a. The coupling constants $\chi_{aa} = eQq_{aa}$, $\chi_{bb} = eQq_{bb}$, and $\chi_{cc} = eQq_{cc}$ are relative to the principal axis system; $q_{aa} = \partial^2 V / \partial a^2$ and so on are the corresponding field gradients. Since Laplace's relation holds,

$$\chi_{aa} + \chi_{bb} + \chi_{cc} = 0, \tag{130}$$

there are only two independent coupling constants. The above may also be applied to an oblate asymmetric rotor with replacement of a by c and b_p by b_0. The selection rules $\Delta J = 0, \pm 1$, $\Delta F = 0, \pm 1$, and $\Delta I = 0$ apply along with those cited previously for $\Delta\tau$. The calculation of the hyperfine structure for an asymmetric top requires a knowledge of the reduced energies and values of $\langle P_a^2 \rangle$ for the two levels involved in the transition. These may be obtained (see Section IV) once the rotational constants are specified. The χ_{aa} and η are then adjusted to obtain agreement between the observed and predicted hyperfine patterns.

As an example of a hyperfine pattern, we give in Fig. 24 the observed spectrum for the $0_{0,0} \rightarrow 1_{1,1}$ transition of CH_3CH_2NC. The splittings due to ^{14}N are very small and high-resolution techniques are required. Since $I(^{14}N) = 1$, we have for the $J = 0$ level, $F = 1$, and for the $J = 1$ level, $F = 0, 1, 2$ and, therefore, only three hyperfine components are possible ($\Delta F = 0$ or ± 1) as observed.

Many molecules have more than one quadrupolar nucleus and the hyperfine structure is made up of many more transitions, and the corresponding coupling pattern can be quite complicated. Theoretical treatments are available, and with improvements in resolution, more of these complicated spectra are being analyzed. As a comparison with Fig. 24, we show the hyperfine coupling pattern exhibited by the $0_{0,0} \rightarrow 1_{0,1}$ transition of HN_3 with

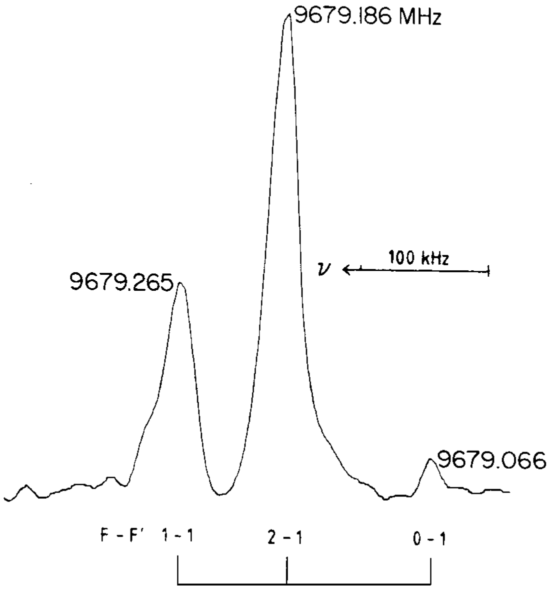


FIGURE 24 The nitrogen quadrupole coupling hyperfine structure for the $0_{0,0} \rightarrow 1_{1,1}$ transition of the asymmetric top ethyl isocyanide, CH_3CH_2NC. The high resolution required to resolve these closely spaced components was obtained by use of Fourier transform microwave spectroscopy. A range of 400 kHz out of a 10 MHz scan is shown. [After Fliege, E., and Dreizler, H. (1984). *Z. Naturforsch.* **40a,** 43.]

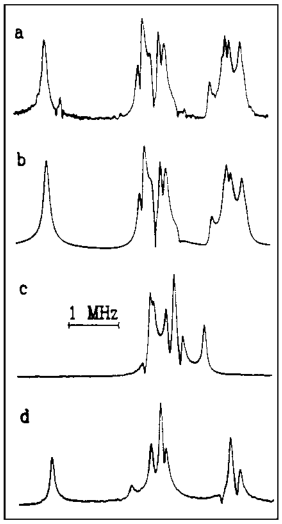


FIGURE 25 The ^{14}N-quadrupole hyperfine structure of the $0_{0,0} \rightarrow 1_{0,1}$ transition of various isotopic species of HN_3. The center frequencies have been aligned for display purposes. The spectra were obtained with a Fourier transform microwave spectrometer. (a) Observed coupling pattern for $H^{14}N_3$, (b) calculated coupling pattern for $H^{14}N_3$, (c) observed coupling pattern for $H^{15}N^{14}N^{14}N$, (d) observed coupling pattern for $H^{14}N^{14}N^{15}N$. [From Gerry, M. C. L., Heineking, N., Mäder, H., and Dreizler, H. (1989). *Z. Naturforsch.* **44a,** 1079.]

three ^{14}N-coupling nuclei. The observed spectra for HN_3, $H^{15}N^{14}N^{14}N$, and $H^{14}N^{14}N^{15}N$ are depicted in Fig. 25a, c, d, respectively. The complexity over Fig. 24 is apparent. The calculated spectrum for HN_3 (Fig. 25b) compares quite well with the observed result (Fig. 25a). For ^{15}N, $I = \frac{1}{2}$ and the hyperfine coupling pattern will depend on only two ^{14}N nuclei. The simplification in the coupling patterns when one ^{14}N is replaced by ^{15}N is apparent from the figure. It follows that for ^{14}N coupling, the hyperfine structure can be simplified by judicious choice of isotopic substitutions.

With improved experimental techniques, it has recently been possible readily to obtain the off-diagnoal elements of the quadrupole coupling tensor χ_{ab}, χ_{bc}, χ_{ac}. However, from one isotopic species only the magnitudes, but not the sign, of the individual off-diagonal elements can be determined along with the sign of the product $(\chi_{ab} \cdot \chi_{bc} \cdot \chi_{ac})$. For the case of one off-diagonal element, the sign is also undetermined. This ambiguity can be removed, however, by use of isotopic information (other than the coupling atom). *Ab initio* calculations of the coupling constants also can be helpful in clarifying the sign of $\chi_{\alpha\beta}$.

For purposes of interpreting the coupling constants in terms of chemical bonding, the constants χ_{aa}, χ_{bb}, and χ_{cc} are usually transformed by a similarity transformation, $\tilde{\mathbf{R}}\chi\mathbf{R}$, to an axis system along and perpendicular to the bond axis involving the coupling nucleus. These elements are often designated by χ_x, χ_y, χ_z. On the other

hand, for linear and symmetric rotors, the field gradient is symmetric ($\chi_{xx} = \chi_{yy}$) about the bond axis z, which corresponds to a principal axis, and from Eq. (130) it follows that $\chi_x = \chi_y = -\frac{1}{2}x_z = -\frac{1}{2}\chi$.

C. Quadrupole Coupling Constants and the Chemical Bond

Considerable information on chemical bonding has been obtained from the quadrupole coupling constants. The basic approach is to relate the nuclear coupling in molecules to the known coupling in free atoms. The field gradient in atoms is primarily determined by the p electrons, and in a molecule any factor that affects the p electron distribution gives rise to a change in the atomic coupling constant. Relations between the charge distribution near the coupling nucleus and the coupling constants are available. These analyses have provided additional insights into the concepts of ionic character, hybridization, π-bonding, and electronegativities. Table XXVII illustrates the dependence of the field gradient q on the electronic environment, showing eQq for the Cl nucleus for a series of molecules. The value for FCl is larger than for atomic Cl, which indicates a positive charge on the Cl atom, F^-Cl^+. At the other extreme, for the highly ionic bond in K^+Cl^-, where the Cl nucleus is surrounded by a symmetric charge distribution, we find $q \equiv 0$. On the other hand, the coupling constant for BrCl is near that for atomic chlorine, which indicates essentially a covalent bond with little ionic character. As one proceeds from ICl to AlCl, the bond becomes less covalent and more ionic in character.

D. Magnetic Interactions

Although we shall not treat magnetic interactions, the reader should be aware that small magnetic hyperfine splittings are possible in singlet electronic ground states and have been observed. This arises from the interaction ($\mathcal{H} = -\boldsymbol{\mu}_I \cdot \mathbf{H}$) of the nuclear spin magnetic moment $\boldsymbol{\mu}_I$ and a weak magnetic field **H** generated by rotation of the molecule. Hence, H is proportional to the angular momentum. High-resolution techniques are required to observe the small magnetic splittings required for evaluation of the nuclear magnetic coupling constants. The theory for combined nuclear quadrupole and magnetic hyperfine interactions has been developed. For CH_2O, the observed hyperfine structure arises solely from the magnetic moments of the two protons, since neither hydrogen nor oxygen has a quadrupole moment. On the other hand, for CHDO there is a quadrupole interaction from D which has a quadrupole moment as well as a magnetic interaction from the nuclear spins of H and D.

TABLE XXVII Comparison of *eQq* (in MHz) for a Chlorine (^{35}Cl) Nucleus in Various Molecules

Molecule	*eQq*
ClF	−145.9
Cl (atomic)	−109.74
BrCl	−102.4
ICl	−85.8
ClCN	−83.4
HC≡CCl	−79.7
CH_3Cl	−74.8
SiH_3Cl	−40.0
AgCl	−37.3
TlCl	−15.8
AlCl	−8.8
KCl	0

XI. ADVANCED EXPERIMENTAL METHODS

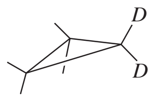
Continued developments in experimental and theoretical techniques have ensured that microwave spectroscopy remains a very active field. Since its inception shortly after World War II, numerous experimental improvements and technological advances have been made. These and various theoretical improvements are discussed and well documented in other articles and books (see Bibliography). Here we highlight a few of the more recent developments.

A. Spectroscopy of Isotope-Substituted Nonpolar Molecules

It was often observed that microwave spectroscopy could not be applied to nonpolar molecules. This observation must be somewhat clarified. Improvements in detection sensitivity have opened the way to measurements on very slightly polar molecules with electric dipole moments approaching 10^{-5} D. Moreover, mechanisms for the generation of such small moments have now been recognized and exploited.

Small dipole moments can be induced in nonpolar molecules by three mechanisms. First, small dipole moments can be vibrationally induced when, for example, symmetric or spherical tops are in excited, degenerate vibrational states. Second, for nonlinear molecules without a center of symmetry, centrifugal distortion can distort the molecule, hence generating a small dipole moment. This mechanism, as discussed in Section IV. B, has now allowed the observation of rotational transitions in the ground vibrational state of spherical tops. Third, nonsymmetric isotopic substitution of a nonpolar molecule can give rise to a small molecular dipole moment by vibrational averaging effects. This case is the most general and is applicable to a large number of nonpolar molecules.

TABLE XXVIII Microwave Studies of Isotopically Induced Polar Molecules

Molecule	Reference
$^{17}O{=}C{=}O$,[a] $\mu = 0.0006$ D; $^{18}O{=}C{=}O$[a], $\mu = 0.0011$ D	Gripp, J., Mäder, H., Dreizler, H., and Teffo, J. L. (1995). *J. Mol. Spectrosc.* **172,** 430
 ,[a] $\mu = 0.0115$ D	Bottcher, O., Heineking, N., and Sutter, D. H. (1990). *J. Mol. Spectrosc.* **139,** 236
$D_2C{=}C{=}CH_2$,[b] $\mu = 0.0053$ D	Meyer, V., and Sutter, D. H. (1993). *Z. Naturforsch.* **48a,** 725

[a] Studied with a waveguide MWFT spectrometer.
[b] Studied with a waveguide bridge-type MWFT spectrometer equipped with flat, oversize Stark cells.

Deuterium substitution often provides dipole moments of sufficient magnitude for ready detection, typically 0.01 D. Such isotopic substitution allows the direct measurement by microwave spectroscopy of many fundamental molecules, such as CH_4, CH_3CH_3, $CH_2{=}CH_2$, and $CH{\equiv}CH$. In addition to new structural information, other advantages are possible. Internal rotation information is difficult to extract from symmetric-top spectra. By converting a symmetric top to an asymmetric top (say ($CH_3CH_3 \rightarrow CH_3CHD_2$), effects such as internal rotation can be more readily studied. Heavy-atom substitution also produces a small dipole moment; however, the result of this type of substitution will be much smaller than for H-atom substitution. For example, in $^{16}O^{12}C^{17}O$, the observed dipole was estimated to be only 6×10^{-4} D. Other examples of small induced dipole moments from isotopic substitution are given in Table XXVIII.

B. Unstable and Semistable Species

Many molecular species studied by microwave spectroscopy are unstable to various degrees, and special preparation techniques, absorption cells, and instrumentation methods have been developed for their investigation. These techniques and methods have been applied to the study of radicals, ions, and semistable molecules. The low operating pressure of typical microwave studies helps in minimizing decomposition from wall collisions and intermolecular collisions.

Radicals (see also Section IV.E.2) are very short-lived, reactive species. They are often produced as products of an RF electric discharge. Alternatively, the products of an electric discharge are allowed to react with another substance to produce the desired radical. In these production methods, a continuous flow of radicals is supplied to the microwave absorption cell. The use of glass absorption cells with Teflon windows is particularly useful for such studies. The large volume-to-surface ratio possible for such cells minimizes radical decomposition. A typical cell is illustrated in Fig. 26. This type of cell, or a variant of this cell, can be used to study molecular ions, semistable molecules, and molecules at high temperature.

Although molecular ions have been detected and assigned (see Section IV.E.2), their number is still small, primarily because of the difficultly in producing significant concentrations. A new technique which increases their density by about two orders of magnitude employs a longitudinal magnetic field along the axis of a glow discharge tube. The addition of the magnetic field increases the length of the ion-rich negative glow and the concentration of the molecular ions. The signal enhancement by application of a magnetic field for HCO^+ is illustrated in Fig. 27. The inset depicts the glass pipe absorption cell (5 ft long, 1.5 in. inside diameter). The glass pipe is connected to short transition sections that flare to 4 in. The transition sections house cylindrical electrodes of the same inside diameter as the glass pipe to maximize microwave transmission. The solenoid provides a field of up to 300 G.

By employment of high-temperature cells, numerous molecules have been studied that at room temperature would not have sufficient vapor pressure to give an observable microwave spectrum. Various alkali halides, for

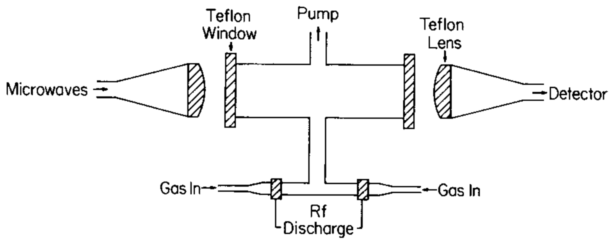


FIGURE 26 Details of a typical free space cell for the study of radicals or other unstable species. The reactive species pass into the cell and are subjected to microwave radiation, and the resulting absorption is detected. The molecular species are continuously replaced by the pumping system.

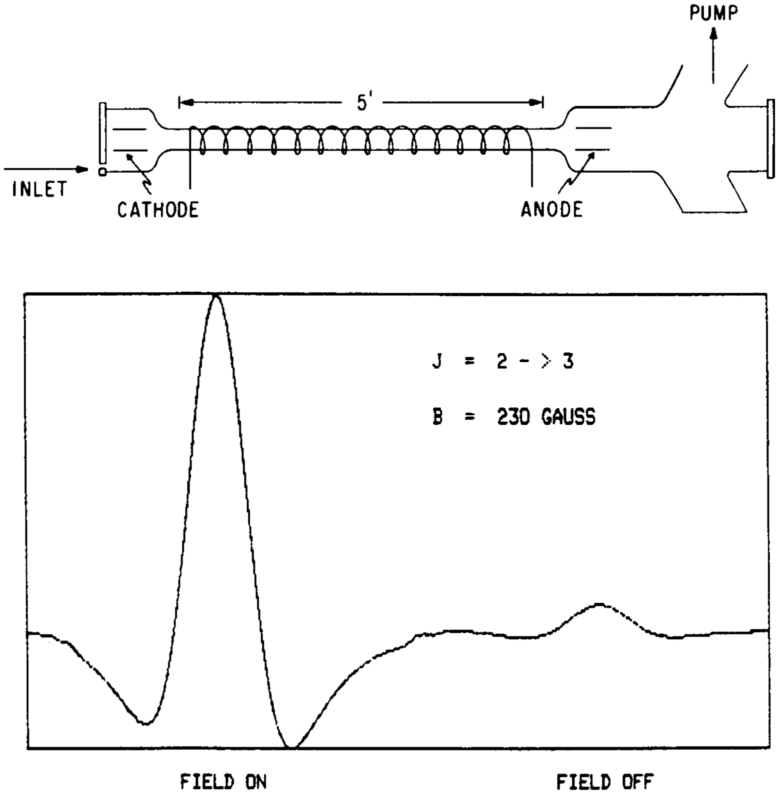


FIGURE 27 Apparatus for enhanced production of positive molecular ions. A 13-mA discharge in a 1:1 mixture of CO and H_2 is used to produce HCO^+. Enhancement of the line at 267.5 GHz is apparent when the field is turned on. [From De Lucia, F. C., Herbst, E., Plummer, G. M., and Blake, G. A. (1983). *J. Chem. Phys.* **78,** 2312.]

example, have been studied this way. *In situ* production in hot absorption cells has been useful for many cases where simple vaporization is not applicable because of dissociation or decomposition. Here, reactions are often made to occur within the hot cell by flowing appropriate constituents into the cell. Likewise, pyrolysis and thermolysis have been developed as useful techniques for producing numerous semistable molecular species.

The lifetimes of semistable molecules are usually on the order of 1 sec; hence, the molecular species are not isolable, but they live long enough to flow through conventional absorption cells at moderate flow rates. Such molecular species are not generally accessible by standard chemical methods and their properties and chemistry are often relatively unknown.

By coupling a reaction flow system with a microwave spectrometer, semistable molecules can be detected and readily identified. A simple production method would involve merely heating an appropriate precursor and passing the resulting decomposition products into an absorption cell. Alternatively, a compound may be heated (or not) and transported with (or without) an inert carrier gas into a reaction zone where reaction with another (heated) substance takes place. The reaction zone may be just before or inside the absorption cell. As general examples, we cite the production of thioketene, $CH_2{=}C{=}S$, by pyrolysis of the trimer $[(CH_3)_2CS]_3$ at 1000°C, preparation of bromoketene, BrHCCO, by pyrolysis of $Br_2HCCOCl$ with zinc metal at 300°C, the synthesis of ClBSe by passing Cl_2Se_2 over boron at 1100°C, and of XNCO (X = Br, I) by passing the halide vapor, X_2, over warm, dry AgNCO. Figure 28 shows an example of a high-temperature reaction flow system. The molecule or molecules to be reacted are passed through a small quartz tube heated to an appropriate temperature. The thermolysis products flow directly into and through the absorption cell of a microwave spectrometer. The flow rate is adjusted to maintain a suitable pressure (ca. 0.1–100 μm Hg). The low operating pressures characteristic of microwave spectroscopy help to reduce unwanted polymerization or decomposition with the metal walls, etc., and the flow technique continually replaces the sample with fresh reaction products.

C. Low-Temperature Spectroscopy

At very low temperatures, the number of thermally populated states is significantly reduced and complex spectra can be greatly simplified. The simplification of complex

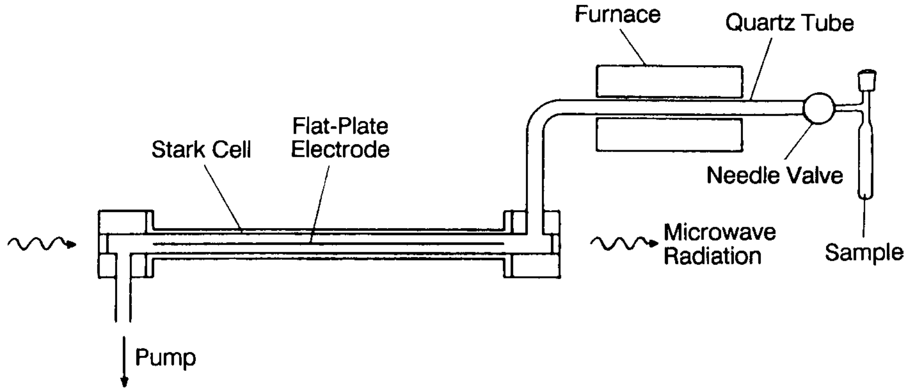


FIGURE 28 Illustration of a high-temperature reaction flow system coupled to a microwave spectrometer for producing short-lived molecules. The gas is made to flow through the Stark cell by means of vacuum system and the flow rate is adjusted by a needle valve. The 8-mm-inside-diameter quartz tube is heated in a 10-cm-long furnace. The distance between the cell and furnace is also about 10 cm. [From Kroto, H. W. (1982)]. *Chem. Soc. Rev.* **11,** 435.]

spectra contributes immensely to the task of assigning the spectra. Moreover, detection is improved because of an increase in the absorption coefficient with decreasing temperature. Temperatures near 1 K have been obtained in microwave studies. Very low temperature spectroscopy provides the opportunity to study weakly bound species, discussed in Section IV.E.3, as well as other processes. Some of the advantages and applications are summarized in Table XXIX.

Very low temperature conditions can be achieved by use of free expansion jets. Numerous weakly bound complexes have been produced by this method (see Sections IV.E.3 and XI.D). Another method is a collisional cooling technique. This technique is particularly versatile; the system is shown in Fig. 29. The system employs a variable-temperature injector to produce an enhanced population in excited vibrational states and liquid nitrogen temperature trapping. The cell is maintained at 77 K and contains a buffer gas of helium or nitrogen, which has a significant vapor pressure at 77 K. A small amount of the gas to be studied is injected into the cooled absorption cell, where it undergoes a series of collisions with the cold background gas as it randomly walks toward the cell wall and freezes out. The spectroscopic gas may be heated by means of the small oven around the injector tube to temperatures up to ca. 1300 K. Since the spectroscopic gas is at low concentration, it does not affect the temperature of the inert buffer gas. The collisions cool the molecule's various degrees of freedom, and typically some 10,000

TABLE XXIX Advantages and Applications of Low-Temperature Microwave Spectroscopy

Improved resolution
Increased sensitivity
Simplification of complex spectra
More direct correspondence between observable parameters and theoretical parameters
Enhance production and lifetime of transient species
Energy transfer
Reaction dynamics at low temperature
Pressure-broadening processes where $h\nu > kT$
Simulate low-temperature and low-pressure environments (outer space)

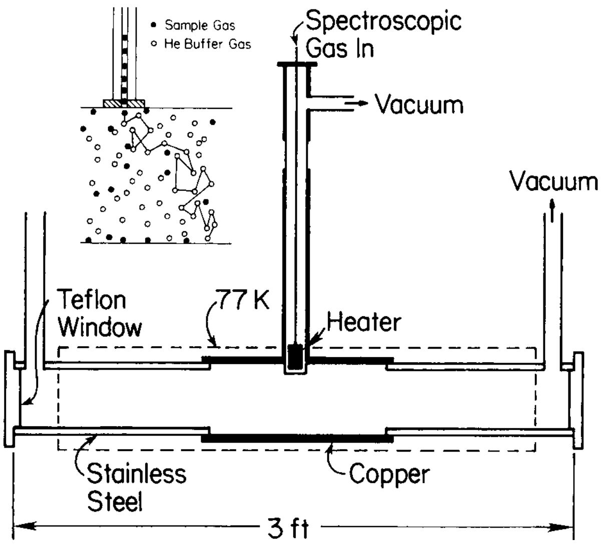


FIGURE 29 Schematic illustration of a collisionally cooled cell. The 4-in.-diameter cell is 3 ft long with the center 1-ft section made of copper and the end sections of stainless steel. The dashed-line region is enclosed in a 77 K medium. Windows for passing microwave radiation are $\frac{1}{2}$-in. Teflon. The injector is a copper tube (0.04-in. inside diameter) surrounded by a small oven at the inlet to the cell. The inset shows a random collision scenario. [After Goyette, T. M., Ebenstein, W. L., and De Lucia, F. C. (1990). *J. Mol. Spectrosc.* **140,** 311.]

collisions occur before the molecule collides with the cell wall. It is found, that in less than 100 collisions, the translational and rotational temperature is cooled to that of the buffer gas. On the other hand, the relaxation of the vibrational degrees of freedom is slower. The vibrational temperature is found to depend on the ejection temperature, the number of collisions with the background gas, and the vibrational relaxation cross section. It is therefore possible to attain separate translational/rotational and vibrational temperatures by control of the injection temperature and the pressure and temperature of the buffer gas.

Collisional Broadening

To extract molecular concentration information and to model the earth's atmosphere accurate knowledge of pressure and temperature effects on spectral line shapes is needed. Such pressure broadening studies are necessary to develop models of the chemistry and physics of the atmosphere and to gain insight into problems associated with air pollution, the greenhouse effect, ozone hole, etc. Spectroscopic remote sensing of planetary and interstellar atmospheres also requires such line broadening information. Line shape information provides direct information on the environment of the molecule, viz., temperature, pressure, collision partner, etc. Principal collision partners for minor atmospheric species are N_2 and O_2. Pressure broadening studies have been carried out at temperatures characteristic of the earth's atmosphere. However, for planetary species, the dominant collision partners are He and H_2. Furthermore, the atmospheres are characterized by low pressures and temperatures, which are difficult to simulate with conventional approaches. At the low temperatures, with conventional techniques, the vapor pressures would be vanishingly small due to condensation. On the other hand, the collisional cooling technique discussed here provides an ideal laboratory method to simulate planetary conditions and similar low-temperature, low-pressure conditions.

In linewidth studies, the width is measured versus pressure and temperature. The linewidth varies with pressure at a given temperature as

$$\Delta\nu = \gamma P + \Delta\nu_0, \tag{131}$$

where P is the pressure and γ is the pressure broadening coefficient. Here $\Delta\nu$ denotes the total linewidth. The temperature dependence of γ is taken as

$$\gamma(T) = \gamma_0(T_0/T)^n, \tag{132}$$

where γ_0 is the coefficient at the reference temperature T_0 and n is a constant. At each temperature, linewidth measurements are made at various pressures. The linewidth $\Delta\nu$ is extracted from the spectral line shape. These data provide γ from Eq. (131) and n from the temperature dependence of Eq. (132).

As an example, typical of pressure broadening studies, we consider the $2_{2,0} \rightarrow 3_{1,3}$ transition of water at 183 GHz with H_2 as a collision partner in the temperature range 80–600 K and pressure range 0.05–1.0 Torr. Above 150 K, the temperature dependence is given by the above relation with $\gamma_{H2}(300\text{ K}) = 3.20 \pm 0.08$ MHz/Torr and $n = 0.95 \pm 0.07$. However, below 150 K, there is a significant deviation from the power law. This result indicates that H_2 as a collision partner is considerably more complex, and H_2 does not act as a classical hard sphere. This is also evident from the value of n since for a hard-sphere model $n = 0.50$. On the other hand, for He as a collision partner, n over the whole temperature range is found to be 0.49 ± 0.02, indicating that He behaves essentially as a classical hard sphere in collisions with H_2O.

Similarly, the proper interpretation of radio astronomical spectral lines from dense interstellar clouds requires collisional information involving ions at low temperatures. By combining the considerations reflected in Figs. 27 and 29, it is possible to obtain the pressure broadening of molecular ions at very low temperatures. This has been demonstrated for the ion HCO^+ and the collision partner H_2. Likewise, extension of direct time-resolved measurements, as discussed in Section IV.D, has been recently carried out incorporating the collisional cooling technique discussed here.

D. Fourier-Transform Microwave Spectroscopy

The technique of Fourier-transform microwave spectroscopy (FTMS) has been applied to the study of a number of weakly bonded complexes, the observation of weak isotopic species, and the resolution of hyperfine structure. It is characterized by higher resolution and sensitivity than conventional Stark-modulated spectroscopy. The superior resolution is demonstrated in Figs. 24 and 25.

In this method, a short, intense microwave pulse is applied to the sample. This pulsed microwave excitation of the sample produces a transient emission signal which is detected by a transient signal averager. The time response of the system is hence observed. Both waveguide-based sample cells and cavity-based cells have been employed. To produce significant transient emission, a high power source (order of watts) is required for the waveguide-based system. Lower power sources are applicable to the cavity systems because of the very narrow bandwidth of such systems. The transient emission is usually averaged over many cycles to improve the signal-to-noise ratio. This

time-domain signal is converted to the frequency domain by calculating the Fourier transform. Earlier waveguide MWFT spectrometers developed by Drizler and coworkers used square cross-section sample cells. Improved versions have used circular waveguide sample cells and bridge-type arrangements where one arm of the bridge is the sample cell and the other arm the reference cell.

In the study of complexes, a Fabry-Perot cavity is employed as a sample cell along with a pulsed supersonic nozzle source. The weakly bonded complex is produced at very low temperatures (a few degrees Kelvin), by expanding (adiabatically) as a short pulse a mixture of the reactive species in an inert buffer gas. Weak complex formation occurs undoubtedly via three-body collisions, for example,

$$\mathrm{Xe} + \mathrm{Xe} + \mathrm{HCl} \rightarrow \mathrm{Xe}\cdot\cdot\mathrm{HCl} + \mathrm{Xe},$$

and the low temperatures obtained help stabilize the complex.

A simplified diagram is given in Fig. 30 of a pulsed nozzle FTMS which has been developed by Flygare and coworkers. The cavity consists of two spherical concave metal mirrors. Each mirror has a centrally located small hole (iris) to couple the radiation into and out of the cavity with a waveguide. Before exiting, radiation reflects back and forth between the mirrors, giving the cavity a long effective path length and high sensitivity. The cavity resonant frequency, determined by the mirror dimensions, is varied by very small movement of one of the mirrors. In operation, a valve is opened and a short pulse (a few milliseconds) of gas is introduced into the cavity and shortly thereafter a pulse (5 μsec) of microwave radiation ν from the BWO source is produced by opening and closing pin diode switch 1. Switch 2 is then opened and the resulting coherent emission is detected. The signal is processed and digitized, and the Fourier transform (frequency-domain

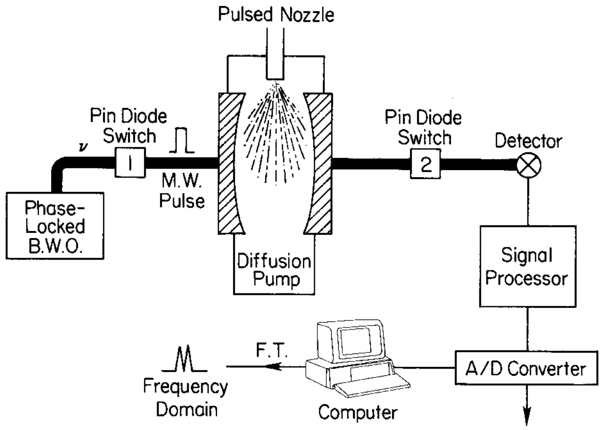


FIGURE 30 A simplified pulsed nozzle Fourier transform microwave spectrometer.

signal) computed from the time-domain signal to give the spectrum. Because the cavity is tuned to ν and has a very narrow bandwidth ($\sim$1 MHz), a high radiation density is obtained. This produces appreciable polarization of the complex when rotational transitions lie inside the frequency bandwidth. Switch 2 serves to protect the detector, and its opening is delayed to allow the microwave pulse to dissipate within the cavity before detection of the weak, longer duration, coherent resonant rotational emission. Gas pulses are repeated by rapid evacuation of the cavity at a rate of $\sim$1 per second. A timing scheme is set up to properly sequence the various events and to allow alternate collection of the background which is subtracted from the molecular emission. Various cycles are averaged to produce the desired spectrum.

Though the resolution is exceptionally higher, because of the narrow bandwidth, only small elements of the spectra can be covered at a time. Also, the low gas temperature allows primarily the ground vibrational state to be studied.

As the technique has become more mature, various molecular species have been studied, and various techniques have been introduced to increase the versatility of the FT spectrometers. For the cavity-pulsed FTMW spectrometer, in addition to studies of weakly bound clusters, various studies have been carried out on reactive short-lived unstable molecules, as well as stable molecules with low concentrations. In the case of large, stable molecules, for example, the spectral density can be very high because of the small rotational constants and hence closely spaced energy levels. This, coupled with vibrational satellites from low-frequency vibrations and/or large nuclear hyperfine structure, makes the assignment of such spectra difficult. The use of a supersonic expansion greatly simplifies the spectrum by forcing the molecules into the lowest rotational states (low J) and the ground vibrational state. This, coupled with the high sensitivity of the technique, allows the assignment of large molecules or low-concentration molecules.

The rapid expansion process provides for molecular formation in a very short time ($\sim$10 μsec) and leaves molecules in an essentially frozen state where additional chemical reactions via collisions cannot occur. This provides the ability to study unstable molecular clusters, short-lived transient species, etc.

Different strategies can be used to produce the desired molecular species to be studied. For example:

- Two molecules in a buffer gas can be mixed followed by expansion to produce a complex, for example, $\mathrm{Kr}\cdot\cdot\mathrm{HCN}$.
- Rapid mixture of two reactive components, one in a buffer gas, just prior to expansion can be used to produce a complex, for example, $\mathrm{H_3N}\cdot\cdot\mathrm{F_2}$.

- A high-temperature (ca. 1000 K) capillary tube can be used to pyrolyze a molecule in a buffer gas followed by gas expansion
- A dc glow discharge of an appropriate gas mixture can be used to produce the desired molecule. Alternatively, the electric discharge can be used to excite vibrational states in one component of a gas mixture and these excited molecules form a vibrationally excited complex, for example, Ar· · ·HCl (10°0 state).
- Laser ablation of solids can be used to provide metal reactants.

In laser ablation, a metal is vaporized by a high-powered laser (Nd:YAG, 532 nm) and the metal vapor is allowed to react with other species present in a buffer gas. The product is stabilized by expansion through a nozzle into the cavity. This method has been used to produce and study such species as InX (X = F, Cl, Br), AlCl, YCl, YBr, and others.

Besides the perpendicular configuration shown in Fig. 31, it is possible to use a parallel configuration where the nozzle is mounted in one of the mirrors and the molecular beam from the expansion nozzle travels parallel to the axis of the cavity. This configuration, in fact, provides improved resolution.

1. Pulsed-Nozzle MWFT Double-Resonance Spectroscopy

With the configuration of Fig. 31, it is possible to conveniently perform microwave–millimeter wave double-resonance (DR) experiments. Here, the signal transition is in the microwave region and the pump transition is in the millimeter-wave region. The pump radiation is introduced into the cavity by means of an oversized microwave horn mounted inside the vacuum chamber. The interaction region for the millimeter waves with the microwave radiation is a small cylindrical region above the horn.

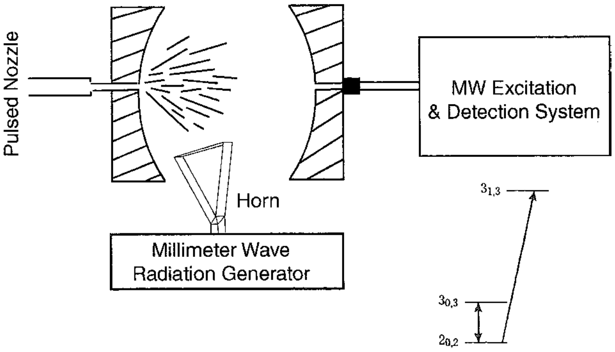


FIGURE 31 A schematic representation of a pulsed cavity FTMW spectrometer with both molecular beam and cavity axis parallel. For DR experiments, a microwave horn is used to inject millimeter-wave radiation into a small interaction zone of the molecular expansion. The inset shows the energy level scheme for a DR experiment. Here the signal frequency was fixed at the $2_{0,2} \rightarrow 3_{0,3}$ transition and this transition was monitored as the pump radiation was scanned. For details of the DR experiment see Jager, W., and Gerry, M. C. L. (1995). *J. Chem. Phys.* **102,** 3587.

The inset in Fig. 31 depicts a typical DR experiment. The double arrow line represent a microwave signal transition and the single arrow a millimeter-wave pump transition. This is only one of four possible double resonance connections for this three-level system. In an actual experiment, the $2_{0,2}$–$3_{0,3}$ microwave transition is monitored while the millimeter-wave pump transition is slowly swept through a given spectral region. A change in the signal transition (DR effect) confirms the energy level structure and provides an accurate measure of the pump transition $2_{0,2}$–$3_{1,3}$. This DR experiment illustrates how spectral assignments may be confirmed. Also, note that the high-frequency pump transition, which lies above the frequency region of the cavity MWFT spectrometer, can be indirectly measured with this DR experiment.

E. Terahertz Region

As apparent from Eq. (1), line intensity increases with increased frequency and hence the terahertz region offers significant advantages for the detection of weak signals. The major methods to obtain high frequencies are (i) frequency multiplication of lower frequency sources and (ii) high-frequency tunable backward-wave oscillators (BWOs). BWOs have been developed that provide radiation to 1.3 THz. These sources coupled with broadband scanning techniques provide capabilities similar to the lower microwave and millimeter-wave frequency regions. Significant improvements in multiplier design employing a planar Schottky diode also allow efficient conversion of fundamental frequencies to over a terahertz. Here, a high-power, lower frequency BWO operating in the range 300–600 GHz is used as the primary source. The region is hence characterized by high sensitivity, high resolution, high measurement accuracy, and broadband tuning capabilities.

SEE ALSO THE FOLLOWING ARTICLES

Analytical Chemistry • Collision-Induced Spectroscopy • Hydrogen Bond • Infrared Spectroscopy • Interstellar Matter • Ion Kinetics and Energetics • Microwave Communications • Millimeter Astronomy • Quantum Chemistry • Time and Frequency

BIBLIOGRAPHY

Aliev, M. R., and Watson, J. K. G. (1985). *In* "Molecular Spectroscopy: Modern Research" (K. N. Rao, ed.), Vol. 3, Academic Press, Orlando, FL.

Baker, J. G. (1979). *In* "Modern Aspects of Microwave Spectroscopy" (G. W. Chantry, ed.), Academic Press, New York.

Carli, B., and Carlotti, M. (1992). "Far-infrared and microwave spectroscopy of the Earth's atmosphere." *In* "Spectroscopy of the Earth's Atmosphere and Interstellar Medium" (K. N. Rao and A. Weber, eds.), Academic Press, San Diego, CA.

Champion, J. P., Loëte, M., and Pierre, G. (1992). "Spherical top spectra." *In* "Spectroscopy of the Earth's Atmosphere and Interstellar Medium" (K. N. Rao and A. Webers, eds.), Academic Press, San Diego, CA.

Cook, R. L., and Jones, G. E. (1974). *In* "Systematic Materials Analysis" (J. H. Richardson and R. V. Peterson, eds.), Academic Press, New York.

De Lucia, F. C. (1976). *In* "Molecular Spectroscopy: Modern Research" (K. N. Rao and C. W. Mathews, eds.), Vol. 2, Academic Press, New York.

Gordy, W., and Cook, R. L. (1984). "Microwave Molecular Spectra," 3rd ed., Wiley-Interscience, New York.

Hirota, E. (1985). *In* "Molecular Spectroscopy: Modern Research" (K. N. Rao, ed.), Vol. 3, Academic Press, Orlando, FL.

Jones, H. (1979). *In* "Modern Aspects of Microwave Spectroscopy" (G. W. Chantry, ed.), Academic Press, New York.

Kroto, H. W. (1975). "Molecular Rotation Spectra," Wiley, London.

Krupnov, A. F. (1979). *In* "Modern Aspects of Microwave Spectroscopy" (G. W. Chantry, ed.), Academic Press, New York.

Legon, A. C. (1983). *Annu. Rev. Phys. Chem.* **34,** 275.

Novick, S. E., Leopold, K. R., and Klemperer, W. (1990). "The structures of weakly bound complexes as elucidated by microwave and infrared spectroscopy." *In* "Atomic and Molecular Clusters" (E. R. Bernstein, ed.), Elsevier, New York.

Mizushima, M. (1975). "The Theory of Rotating Diatomic Molecules," Wiley, New York.

Papousek, D., and Aliev, M. R. (1982). "Molecular Vibrational–Rotational Spectra," Elsevier, New York.

Saykally, R. J., and Woods, R. C. (1981). *Annu. Rev. Phys. Chem.* **32,** 403.

Schmalz, T. G., and Flygare, W. H. (1978). *In* "Laser and Coherence Spectroscopy" (J. I. Steinfeld, ed.), Plenum Press, New York.

Winnewisser, G., Churchwell, E., and Walmsley, C. M. (1979). *In* "Modern Aspects of Microwave Spectroscopy" (G. W. Chantry, ed.), Academic Press, New York.

Winnewisser, G., Herbst, E., and Ungerechts, H. (1992). "Spectroscopy among the stars." *In* "Spectroscopy of the Earth's Atmosphere and Interstellar Medium" (K. Rao and A. Weber, eds.), Academic Press, San Diego, CA.

Nuclear Magnetic Resonance (NMR)

Bernard C. Gerstein
Iowa State University

I. Introduction
II. The Nucleus as a Probe of Molecular Structure; Internal Interactions and the Effects of Motion
III. NMR Spectra of Liquids
IV. NMR in Solids
V. Pulse and Fourier Transform NMR
VI. Two-Dimensional NMR
VII. Multiple Quantum NMR
VIII. Medical Imaging

GLOSSARY

Anisotropic Nonuniform in space; not spherically symmetric.

Chemical shift Shift in NMR frequency due to local nuclear chemical environment as compared to a reference frequency.

Dipolar coupling Interaction between two nuclei behaving as classical magnets.

Isotropic Uniform in all directions in space; spherically symmetric.

Isotropic value One third of the sum of the diagonal elements of a 3×3 matrix representing an ellipsoidal surface.

kHz Thousand of cycles per second (cps).

MHz Millions of cycles per second.

Quadrupole Ellipsoidal electric charge distribution.

THE PHENOMENON of nuclei absorbing resonant radio frequency energy in a static magnetic field is called nuclear magnetic resonance (NMR). This phenomenon is always accompanied by nuclear relaxation. Nuclear magnetic resonance spectroscopy is routinely used to determine structures of molecules in liquids and in organic molecular solids. The applications of NMR spectroscopy are diverse. In medicine, NMR spectroscopy is used to evaluate the extent of damage to heart muscles of patients who have suffered severe coronary infraction. The chemist can use it to examine coal to determine the presence of

carcinogenic molecules. The manufacturer can use it to study the morphology of polymer blends for constructing more durable battery casings.

I. INTRODUCTION

Because some of the terms used to describe the phenomenon of nuclear magnetic resonance are a bit unfamiliar, a more familiar physical example illustrating these terms is presented.

To illustrate the meaning of "resonance," recall that a car driven at a certain speed, generally about 80 km/hr, will sometimes begin to vibrate due to a wheel out of balance. This vibration is an example of the phenomenon of "resonance." In a "resonant energy exchange," at a particular driving frequency (in this case the angular frequency of the wheel), a system with a "resonant state" matching the energy of the driving frequency will begin to absorb energy from an external source in a very efficient manner. The system with the resonant state in this case is the spring suspension system of the car. When the out-of-balance wheel, which is vibrating at all speeds, reaches a vibration frequency matching a "resonant frequency" of the car's suspension system, then there is a "resonant exchange" between the wheel and the remainder of the car, and the entire car begins to vibrate as the energy in the vibrating wheel becomes efficiently transferred to the remainder of the car. The energy transferred in a resonant exchange can be stored, and if not released quickly enough, can serve to destroy the system. In a car, that might mean a tire blowing out from repeated resonant bouncing. In a molecule, it could mean thermal decomposition. Alternatively, the system might relax from its high energy state by releasing energy back to the driving source in an oscillatory fashion, or by releasing energy to its surroundings in some monotonically decreasing manner, perhaps in an exponentially damped fashion. For example, the out-of-balance wheel on a car might simply exchange energy with the suspension system of the car and oscillate indefinitely, if the speed of the car is maintained such that resonant exchange can take place. Another possibility is that the spring suspension system of the car, receiving the vibrations of the out-of-balance wheel, might eventually be shaken apart and fail. A driver of a car with a wheel badly out of balance might automatically change the speed of the car to move away from the resonant situation, and thus allow the vibration to damp away, or relax, by energy interchange between the suspension system of the car and the surroundings; the rapid vibration might have heated the springs, and the heat could have been dissipated in the passing air.

Nuclei possessing a magnetic moment, and placed in a static magnetic field of from 5 to 10 T (for comparison, the earth's field is about .00005 T, or one-half Gauss) may be thought of as developing discrete, or quantized alignments of their magnetic moments with respect to the static magnetic field [i.e., the nuclei develop discrete energies, or states, the energy differences of which are in the radio-frequency (millions of cycles per second, or megahertz) region]. One may relate a frequency ν in cycles per second, to an energy difference ΔE by the relation $\Delta E = h\nu$, where h is Planck's constant, $h = 6.627 \times 10^{-27}$ erg sec. When such nuclei are exposed to radio-frequency radiation at the appropriate resonant driving frequency matching the energy *differences* of nuclear states in a magnetic field, then in much the same manner that the suspension system of a car will efficiently absorb energy from a vibrating wheel, the nuclei undergo resonant absorption of the radio-frequency energy from a driving source, which is a resonant alternating current circuit of the type used in radio broadcasting. Such excitation disturbs the nuclei from their initial equilibrium states in the static field. Changing the driving frequency or the external magnetic field will remove the resonant condition, and the nuclei will cease absorbing energy. Following, and during, a resonant absorption of energy, the nuclei can relax from their excited states back to states of energy intermediate between their initial states before resonant excitation and their excited states.

Two common relaxation processes for nuclei are termed "spin–lattice" or longitudinal relaxation, and "spin–spin" or transverse relaxation. In longitudinal relaxation, the nuclei lose their energy to the molecular framework in which they reside, and relax to their initial state, which may be thought of as alignment of their magnetic moments along the static magnetic field. In transverse relaxation, the nuclei relax to a state in which the net magnetic moment perpendicular to the static field is created by many nuclei having their spins pointing "in phase" in the same direction in the transverse plane, and the nuclei lose their phase coherence, and thus their net nuclear magnetism perpendicular to the static field. In the simplest cases (e.g., that of nuclei in molecules in many liquids), these relaxation processes are described by simple exponential forms. The time constant for longitudinal relaxation is designated by the symbol T_1. After a resonant excitation, the recovery of the nuclear magnetization parallel to the static field is proportional to $(1 - e^{-t/T_1})$. Note that this expression is zero at time $= t = 0$ (i.e., immediately following the excitation), and smoothly reaches unity at times longer than $5T_1$. Similarly, the time constant for relaxation of magnetization transverse to the magnetic field is called T_2, and transverse relaxation is exponential in T_2.

The phenomenon of nuclei absorbing resonant radio-frequency energy in a static magnetic field is called nuclear magnetic resonance (NMR), and this phenomenon is always accompanied by nuclear relaxation. The details of resonant absorption of radio-frequency energy by nuclei in a magnetic field leads to the "spectrum" of absorption lines, and the relaxation processes are responsible for line

intensities and the widths of the observed spectral lines. The study of spectra is called "spectroscopy," hence the term NMR spectroscopy. It is seen, therefore, that at the very least, an NMR experiment requires (1) a source of radio-frequency (rf) oscillation at some radial frequency ω (rad sec^{-1}) tuned to energy splittings between nuclear magnetic states, (2) a magnet developing a static field B, to produce the split nuclear magnetic states, (3) nuclei with magnetic moments M placed in a resonant radio-frequency (RF) circuit to absorb the rf energy at frequency ω, and (4) some means of detecting this energy absorption. An NMR spectrometer is basically a high quality FM radio station and accompanying FM receiver. The carrier is in the (video) megahertz region, generally between 5 and 700 MHz. The information content generally comes through in the (audio) kilohertz region, but the sounds that are produced by resonating nuclei, when sent over an audio speaker, are generally fairly monotonic, and are not nearly as pleasant as those designed by a Mozart. The magnet is an expensive addendum to make a portion of the experiment possible.

The fundamental relation between the experimenter-supplied parameters, B and ω, and the nuclear moment M is

$$\omega \propto BM.$$

The resonant frequency of absorption of energy of magnetic nuclei in a magnetic field is proportional to both the strength of the field, and to the magnetic moment of the nucleus. The resonant condition for NMR may be achieved by varying either B or the driving frequency. As alluded to previously, the local electronic and nuclear environment about a nucleus in a molecule, along with the external magnetic field created by a magnet, contributes to the effective value of B. Thus the resonant NMR frequency is a fingerprint of the local electronic environment of the nucleus, but depends upon the external magnetic field, which is at the control of the experimenter. The magnetic moment of a nucleus is a quantity fixed by nature, and is not an experimental variable. Table I lists all of the known magnetic nuclides, their resonant frequencies in the absence of interactions associated with the atomic or molecular environment at a field in which ^{1}H resonates at 100 MHz, and relavent added material, which will become more meaningful as further information is developed. Note that from a quick glance at Table I, it is possible to infer that the physician, the materials scientist, the chemist, the physicist, the polymer chemist, the solid state scientist, the geologist, and the engineer all have problems that may be attacked with the help of NMR, since workers in all of these specialties deal with systems containing one or more of the nuclei listed.

Table I indicates that each magnetic nucleus has a number of fingerprints. One is its "nuclear spin quantum number" I, which is proportional to its magnetic moment M, the proportionality factor being the gyromagnetic ratio γ; $M = \gamma I$. The values of γ, and therefore the resonant frequencies for NMR at fixed field differ for each nuclide. For example, ^{1}H and ^{13}C have spin quantum numbers $I = 1/2$, whereas ^{6}Li and ^{174}Lu are spin 1. ^{27}Al is spin 5/2. This spin quantum number imparts a special character to the nucleus' ability to detect its local molecular architecture, as will be seen in Section II. The NMR absorption spectra, examples of which are shown in Sections III and IV, are generally represented on an intensity (ordinate) vs frequency (abscissa) plot, and appear as a series of peaks of various widths and shapes that are a reflection of the local molecular environment of the nuclei under observation. This is to say that the local environments of nuclei in matter supply effective fields, B_{eff}, which may be used to infer that environment.

We now inquire in more detail as to why NMR has this remarkable capability, and why this resonant spectroscopy is such a powerful tool, relative to other spectroscopies, such as ultraviolet and infrared spectroscopies.

II. THE NUCLEUS AS A PROBE OF MOLECULAR STRUCTURE; INTERNAL INTERACTIONS AND THE EFFECTS OF MOTION

A nucleus residing in a molecule, either in a solid or a liquid sample, has access to quite an intimate view of its local molecular architecture. This nucleus senses the locations and types of its nearest neighbors, and in a diffuse manner, the bulk matter around it. In addition, this nucleus is sensitive to motion of its environment. The nucleus, when properly interrogated with resonant excitations, can give detailed information about its local molecular environment when that environment is motionless. In addition, the alteration of this information caused by molecular motion is used to infer details of such motion. It is this type of information which, when properly interpreted as indicated in the introduction, can lead to the wide variety of applications described there.

The sensitivity of the nucleus to its environment and to motion are all the result of the arrangements of molecular framework electrons and nuclei about the nucleus in question. The effects of this molecular framework upon the effective magnetic fields, and thus upon the resonance frequencies of nuclei in matter are generally separated into four contributions, termed interactions: these are designated (1) "shielding," (2) "dipolar coupling," (3) electric field gradients, or "quadrupolar coupling," and (4) "scalar coupling." These interactions are all anisotropic. This means that they are directionally dependent on the relative orientations of the static magnetic

TABLE I Properties of Magnetically Active Nuclides[a,b]

Atomic weight/ element	Spin	Natural abundance (%)	Sensitivity (versus ^{13}C)	Quadrupole moment (10^{-28} M^2)	Gyromagnetic ratio (10^7 rad T^{-1} sec^{-1})	Resonance frequency (^{1}H TMS 100 MHz)
^{1}H	$\frac{1}{2}$	99.985	5.68×10^3	—	26.7510	100.0000
^{2}H	1	0.015	8.2×10^{-3}	2.73×10^{-3}	4.1064	15.351
^{3}H	$\frac{1}{2}$	—	—	—	28.5335	160.663
^{3}He	$\frac{1}{2}$	0.00014	3.26×10^{-3}	—	−20.378	76.178
^{6}Li	1	7.42	3.58	-8×10^{-4}	3.9366	14.716
^{7}Li	$\frac{3}{2}$	92.58	1.54×10^3	-4.5×10^{-2}	10.3964	38.864
^{9}Be	$\frac{3}{2}$	100	78.8	5.2×10^{-2}	−3.759	14.052
^{10}B	3	19.58	22.1	7.4×10^{-2}	2.8740	10.744
^{11}B	$\frac{3}{2}$	80.42	7.54×10^2	3.55×10^{-2}	8.5794	32.072
^{13}C	$\frac{1}{2}$	1.108	1.00	—	6.7283	25.145
^{14}N	1	99.63	5.69	1.6×10^{-2}	1.9331	7.226
^{15}N	$\frac{1}{2}$	0.37	2.19×10^{-2}	—	−2.7116	10.137
^{17}O	$\frac{5}{2}$	0.037	6.11×10^{-2}	-2.6×10^{-2}	−3.6264	13.556
^{19}F	$\frac{1}{2}$	100	4.73×10^3	—	25.181	94.094
^{21}Ne	$\frac{3}{2}$	0.257	3.59×10^{-2}	9×10^{-2}	−2.1118	7.894
^{23}Na	$\frac{3}{2}$	100	5.25×10^2	0.12	7.0761	26.452
^{25}Mg	$\frac{5}{2}$	10.13	1.54	0.22	−1.6375	6.122
^{27}Al	$\frac{5}{2}$	100	1.17×10^3	0.149	6.9704	26.057
^{29}Si	$\frac{1}{2}$	4.70	2.09	—	−5.3146	19.867
^{31}P	$\frac{1}{2}$	100	3.77×10^2	—	10.8289	40.481
^{33}S	$\frac{3}{2}$	0.76	9.73×10^{-2}	-5.5×10^{-2}	2.0534	7.676
^{35}Cl	$\frac{3}{2}$	75.53	20.2	-8.0×10^{-2}	2.6210	9.798
^{37}Cl	$\frac{3}{2}$	24.47	3.8	-6.32×10^{-2}	2.1817	8.156
^{39}K	$\frac{3}{2}$	93.1	2.69	5.5×10^{-2}	1.2483	4.666
^{40}K	4	0.012	3.52×10^{-3}	$(-)^c$	−1.552	5.801
^{41}K	$\frac{3}{2}$	6.88	3.28×10^{-2}	6.7×10^{-2}	0.6851	2.561
^{43}Ca	$\frac{7}{2}$	0.145	5.27×10^{-2}	−0.05	−1.8001	6.729
^{45}Sc	$\frac{7}{2}$	100	1.71×10^3	−0.22	6.4982	24.292
^{47}Ti	$\frac{5}{2}$	7.28	0.864	0.29	±1.5084	5.639
^{49}Ti	$\frac{7}{2}$	5.51	1.18	0.24	±1.5080	5.638
^{50}V	6	0.24	0.755	±0.21	2.6491	9.970
^{51}V	$\frac{7}{2}$	99.76	2.15×10^3	-5.2×10^{-2}	7.0362	26.303
^{53}Cr	$\frac{3}{2}$	9.55	0.49	$\pm 3 \times 10^{-2}$	−1.5120	5.652
^{55}Mn	$\frac{5}{2}$	100	9.94×10^2	0.55	6.6195	24.745
^{57}Fe	$\frac{1}{2}$	2.19	4.2×10^{-3}	—	0.8661	3.238
^{59}Co	$\frac{7}{2}$	100	1.57×10^3	0.40	6.3472	23.727
^{61}Ni	$\frac{3}{2}$	1.19	0.24	0.16	−2.3904	8.936
^{63}Cu	$\frac{3}{2}$	69.09	3.65×10^2	−0.211	7.0965	26.528
^{65}Cu	$\frac{3}{2}$	30.91	2.01×10^2	−0.195	7.6018	28.417
^{67}Zn	$\frac{5}{2}$	4.11	0.665	0.15	1.6737	6.257
^{69}Ga	$\frac{3}{2}$	60.4	2.37×10^2	0.178	6.420	24.001
^{71}Ga	$\frac{3}{2}$	39.6	3.19×10^2	0.112	8.158	30.497
^{73}Ge	$\frac{9}{2}$	7.76	0.617	−0.2	−9.331	3.488
^{75}As	$\frac{3}{2}$	100	1.43×10^2	0.3	4.5804	17.123
^{77}Se	$\frac{1}{2}$	7.58	2.98	—	5.1018	19.072
^{79}Br	$\frac{3}{2}$	50.54	2.26×10^2	0.33	6.7023	25.054

(*continues*)

TABLE I (*Continued*)

Atomic weight/ element	Spin	Natural abundance (%)	Sensitivity (versus ^{13}C)	Quadrupole moment (10^{-28} M^2)	Gyromagnetic ratio (10^7 rad T^{-1} sec^{-1})	Resonance frequency (^{1}H TMS 100 MHz)
^{81}Br	$\frac{3}{2}$	49.46	2.77×10^2	0.28	7.2246	27.007
^{83}Kr	$\frac{9}{2}$	11.55	1.23	0.15	−1.029	3.848
^{85}Rb	$\frac{5}{2}$	72.15	43	0.25	2.5828	9.655
^{87}Rb	$\frac{3}{2}$	27.85	2.77×10^2	0.12	8.7532	32.721
^{87}Sr	$\frac{9}{2}$	7.02	1.07	0.36	−1.1593	4.334
^{89}Y	$\frac{1}{2}$	100	0.668	—	−1.3108	4.900
^{91}Zr	$\frac{5}{2}$	11.23	6.04	−0.21	−2.4868	9.296
^{93}Nb	$\frac{9}{2}$	100	2.740×10^3	−0.2	6.5476	24.476
^{95}Mo	$\frac{5}{2}$	15.72	2.88	±0.12	1.7433	6.517
^{97}Mo	$\frac{5}{2}$	9.46	1.84	±1.1	−1.7799	6.654
^{99}Tc	$\frac{9}{2}$	100	1.562×10^{3}[d]	−0.19[d]	6.0211	22.508
^{99}Ru	$\frac{5}{2}$	12.72	0.83	7.6×10^{-2}	−1.2343	4.614
^{101}Ru	$\frac{5}{2}$	17.07	1.56	0.44	−1.3834	5.171
^{103}Rh	$\frac{1}{2}$	100	0.177	—	−0.8520	3.185
^{105}Pd	$\frac{3}{2}$	22.23	1.41	0.8	−0.756	4.576
^{107}Ag	$\frac{1}{2}$	51.82	0.195	—	−1.0828	4.048
^{109}Ag	$\frac{1}{2}$	48.18	0.276	—	−1.2448	4.654
^{111}Cd	$\frac{1}{2}$	12.75	6.73	—	−5.6714	21.201
^{113}Cd	$\frac{1}{2}$	12.26	7.6	—	−5.9328	22.178
^{113}In	$\frac{9}{2}$	4.28	83.8	1.14	5.8493	21.866
^{115}In	$\frac{9}{2}$	95.72	1.89×10^3	0.83	5.8618	21.913
^{115}Sn	$\frac{1}{2}$	0.35	0.695	—	−8.792	32.86
^{117}Sn	$\frac{1}{2}$	7.61	19.54	—	−9.5319	35.632
^{119}Sn	$\frac{1}{2}$	8.58	25.2	—	−9.9756	37.291
^{121}Sb	$\frac{5}{2}$	57.25	5.20×10^2	−0.53	6.4016	23.931
^{123}Sb	$\frac{7}{2}$	42.75	1.11×10^2	−0.68	3.4668	12.959
^{123}Te	$\frac{1}{2}$	0.89	0.89	—	−7.0006	26.170
^{125}Te	$\frac{1}{2}$	7.0	12.5	—	−8.4398	31.550
^{127}I	$\frac{5}{2}$	100	5.3×10^2	−0.79	5.3525	20.009
^{129}Xe	$\frac{1}{2}$	26.44	31.8	—	−7.4003	27.658
^{131}Xe	$\frac{3}{2}$	21.18	3.31	−0.12	2.1939	8.200
^{133}Cs	$\frac{7}{2}$	100	2.69×10^2	-3×10^{-3}	3.5087	13.116
^{135}Ba	$\frac{3}{2}$	6.59	1.83	0.18	2.6575	9.934
^{137}Ba	$\frac{3}{2}$	11.32	4.41	0.28	2.9728	11.113
^{138}La	5	0.09	0.43	−0.47	3.5295	13.194
^{139}La	$\frac{7}{2}$	99.91	3.36×10^2	0.21	3.7787	14.126
^{141}Pr	$\frac{5}{2}$	100	1.66×10^3	-5.9×10^{-2}	7.836	29.291
^{143}Nd	$\frac{7}{2}$	12.17	2.31	−0.48	1.455	5.438
^{145}Nd	$\frac{7}{2}$	8.3	0.37	−0.25	0.895	3.346
^{147}Sm	$\frac{7}{2}$	14.97	1.26	−0.21	1.104	4.128
^{149}Sm	$\frac{7}{2}$	13.83	0.59	6×10^{-2}	0.880	3.289
^{151}Eu	$\frac{5}{2}$	47.82	4.83×10^2	1.16	6.634	24.801
^{153}Eu	$\frac{5}{2}$	52.18	45.3	2.9	2.930	10.952
^{155}Gd	$\frac{3}{2}$	14.73	0.23	1.6	1.022	3.820
^{157}Gd	$\frac{3}{2}$	15.68	0.48	2	1.277	4.775
^{159}Tb	$\frac{3}{2}$	100	3.31×10^2	1.3	6.067	22.679

(*continues*)

TABLE I (*Continued*)

Atomic weight/ element	Spin	Natural abundance (%)	Sensitivity (versus ^{13}C)	Quadrupole moment (10^{-28} M^2)	Gyromagnetic ratio (10^7 rad T^{-1} sec^{-1})	Resonance frequency (^1H TMS 100 MHz)
^{161}Dy	$\frac{5}{2}$	18.88	0.45	1.4	0.881	3.295
^{163}Dy	$\frac{5}{2}$	24.97	1.59	1.6	1.226	4.584
^{165}Ho	$\frac{7}{2}$	100	1.03×10^3	2.82	5.487	20.513
^{167}Er	$\frac{7}{2}$	22.94	0.66	2.83	0.773	2.890
^{169}Tm	$\frac{1}{2}$	100	3.21	—	−2.21	8.272
^{171}Yb	$\frac{1}{2}$	14.27	4.05	—	4.72	17.612
^{173}Yb	$\frac{5}{2}$	16.08	1.14	(—)[b]	1.31	4.852
^{174}Lu	1	—	—	—	—	—
^{175}Lu	$\frac{7}{2}$	97.41	1.56×10^2	5.68	3.05	11.407
^{176}Lu	7	2.59	5.14	8.1	2.10	7.872
^{177}Hf	$\frac{7}{2}$	18.50	0.88	4.5	0.95	4.008
^{179}Hf	$\frac{9}{2}$	13.75	0.27	5.1	−0.609	2.518
^{181}Ta	$\frac{7}{2}$	99.988	2.04×10^2	3	3.2073	11.990
^{183}W	$\frac{1}{2}$	14.28	5.89×10^{-2}	—	1.1145	4.166
^{185}Re	$\frac{5}{2}$	37.07	2.8×10^2	2.8	6.0255	22.525
^{187}Re	$\frac{5}{2}$	62.93	4.90×10^2	2.6	6.0862	22.752
^{187}Os	$\frac{1}{2}$	1.64	1.14×10^{-3}	—	0.6105	2.282
^{189}Os	$\frac{3}{2}$	16.1	2.13	0.8	2.0773	7.765
^{191}Ir	$\frac{3}{2}$	37.3	2.3×10^{-2}	1.5	0.539	1.718
^{193}Ir	$\frac{3}{2}$	62.7	5.0×10^{-2}	1.4	0.391	1.871
^{195}Pt	$\frac{1}{2}$	33.8	19.1	—	5.7412	21.462
^{197}Au	$\frac{3}{2}$	100	6.0×10^{-2}	0.58	0.357	1.729
^{199}Hg	$\frac{1}{2}$	16.84	5.42	—	4.7912	17.911
^{201}Hg	$\frac{3}{2}$	13.22	1.08	0.5	−1.7686	6.612
^{203}Tl	$\frac{1}{2}$	29.50	2.89×10^2	—	15.3078	57.224
^{205}Tl	$\frac{1}{2}$	70.50	7.69×10^2	—	15.4584	57.787
^{207}Pb	$\frac{1}{2}$	22.6	11.8	—	5.5797	20.858
^{209}Bi	$\frac{9}{2}$	100	7.77×10^2	−0.4	4.2986	16.069
^{209}Po	$\frac{1}{2}$	—	—	—	—	—
^{235}U	$\frac{7}{2}$	0.72	4.9×10^{-3}	4.1	0.479	1.791

[a] Most values taken from Brevard, C., and Grager, P. (1981). "Handbook of High Resolution Multinuclear NMR," Wiley (Interscience), New York, pp. 80–211.

[b] Some values taken from the Bruker NMR-NQR Periodic Table; Harris, R. K., and Mann, B. E. (1978). "NMR and the Periodic Table," Academic Press, London, pp. 5–7; Pople, J. A., Schneider, W. G., and Bernstein, H. J. (1959). "High-Resolution Nuclear Magnetic Resonance," McGraw-Hill, New York, pp. 480–485; Harris, R. K., private communication.

[c] Poorly known or unknown.

[d] Franklin, K. J., Lock, C. J. L., Sayer, B. G., and Schrobilgen, G. J. (1982). *J. Am. Chem. Soc.* **104,** 5303–5306.

field, and the three-dimensional coordinate system orienting the particular interactions. A nucleus that experiences all of the above four effects of the molecular framework will, in general, have a set of resonance frequencies that are a reflection of all of these contributions. The physical origins of each of these are now discussed in turn, and for simplicity, the effect on the resonance frequency due to each of the four contributions from the molecular framework is discussed as if that were the only contribution present. The observed spectrum of many nuclei is effectively due to a single one of the above contributions, so it makes sense to discuss them one at a time.

A. Shielding; Chemical Shifts

NMR was discovered in 1945 independently by Bloch and co-workers and by Purcell and co-workers. It was

not widely used by chemists, however, until the discovery, five years later by Procter and Yu, that a given nuclear type in different chemical environments (e.g., ^{14}N in the ammonium ion and in the nitrate ion in an aqueous solution of ammonium nitrate, NH_4NO_3) would absorb radio-frequency irradiation at a frequency specific to the chemical environment; nitrogen in the ammonium ion absorbed at a frequency different than nitrogen in the nitrate ion. Subsequent to this discovery, NMR was rapidly assimilated into the chemical laboratory for routine determination of structures of molecules in liquids.

For many nuclei in liquids, the frequency differences of a given nucleus, associated with differences in chemical functionality of the atom in which the nucleus resides, or shifts, known as *chemical shifts* from some standard, are of order of parts per million (ppm) of the applied field. For example, a proton in a methyl group and a proton in an OH group in methyl alcohol, CH_3OH, will have a frequency difference of roughly 7 ppm. Thus, at a resonant frequency of 300 MHz for the proton in the methyl group, the proton in the OH group will absorb at 300×10^6 $(1\text{–}7 \times 10^{-6})$ Hz, or the *difference* in resonance frequencies between ^1H in $CH_3–$ and ^1H in —OH will be 1,800 Hz; this is the desired information in the audio region of frequencies. Currently, the differences in resolution of absorption lines of nuclei in different environments (to be clearly distinguished from quantitative detection, i.e., amount) available using the highest field magnets now produced, which currently are 14 T, and liquid samples, is equivalent to being able to distinguish two meter sticks standing half a meter apart on the moon, when observing from the earth. One of the most powerful fingerprints of nuclei available to the practicing chemist, the chemical shift permits both a quantitative and qualitative analysis of the molecules containing the nucleus under investigation, since individual chemical functionalities such as hydrogen in —CH_3, and hydrogen in —OH can readily be distinguished, and the intensity of the NMR lines corresponding to hydrogen in these two different environments is proportional to the number of hydrogens in that environment. For example, the NMR spectrum of protons in pure methyl alcohol would consist of two lines about 7 ppm apart, with an intensity ratio $I(CH_3)/I(OH) = 3{:}1$. Further, protons in all methyl groups resonate in a small frequency range compared to the difference between protons in CH_3 and protons in OH. A similar statement applies to protons in all hydroxyl groups, with some understandable exceptions, so one talks of the "methyl group range of absorption," etc.

The origin of this "chemical shift" is that a nucleus in a molecule (including infinite molecules such as metals) is surrounded by an electron charge cloud that is a reflection of the chemical bonding about the nucleus, and that is in general some complicated shape. This is to say that it is not a spherical cloud, as it would be, for example, for a nucleus in an inert gas such as helium. When this cloud is exposed to a static magnetic field, it responds in accord with the laws of quantum mechanics by effectively developing a field that algebraically adds to the static field. In other words, the nucleus is "shielded" from the external field by the molecular electronic cloud about it. This shielding interaction causes the magnetic field that is seen by the nuclei in the molecule to be different than the static field due to the magnet in the NMR experiment. As mentioned previously, the basic relation in the NMR experiment is $\omega = MB$. This relation, taking into account the fact that the effective magnetic field at the nucleus is provided both by the static, external field B_0 supplied by the experimenter and by the local magnetic field due to the environment of the nucleus, now becomes $\omega \propto MB_{\text{eff}}$. Now, however, the product MB_{eff} depends upon the direction of the external magnetic field relative to the nonspherical electronic environment producing the shielding shift. We, therefore, digress a moment to consider a physical picture of an anisotropic shielding environment that will be useful in our discussion of all of the four effects of the molecular framework upon resonance frequencies of nuclei in NMR experiments.

A useful pictorial representation of an anisotropic shielding environment is provided by an ellipsoid with three unequal axes, shown in Fig. 1. The relation between the frequency observed in an NMR experiment, ω_{obs}, and this ellipsoid is as follows: If the ellipsoid represents the anisotropic shielding, or chemical shift interaction, the observed angular resonance frequency when the external magnetic field B_0 is parallel to the x axis of the ellipsoid would be given by the simple equation

$$\omega_{\text{obs}} = \gamma B_{\text{eff}} = \gamma B_0(1 - \sigma_{xx}).$$

Here, σ_{xx} is the magnitude of the x axis of the shielding ellipsoid. Clearly, γB_0 represents the NMR frequency in

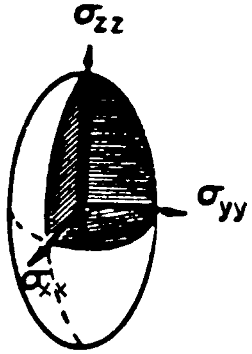


FIGURE 1 Representation of the anisotropy of an internal interaction as an ellipsoid. The principal axes of the ellipsoid represent resonant frequencies for absorption. The three angles orienting this ellipsoid with an arbitrary coordinate system represent the other three independent pieces of information.

the absence of shielding, and the term $-\gamma B_0\sigma_{xx}$ is the frequency shift due to the anisotropic shielding along the x axis of the shielding ellipsoid. The observed resonance frequencies with the external field parallel to the y and z axes of the shielding ellipsoid would be correspondingly different if the values of σ_{yy}, and σ_{zz} differ from σ_{xx}. With the convention $\sigma_{xx} > \sigma_{yy} > \sigma_{zz}$, and B_0 in some direction other than parallel to one of these axes, the observed resonance frequency for a given nucleus will lie between that with the field parallel to the x, and that with the field parallel to the z axes of the shielding ellipsoid. The spectral width (i.e., the range of resonance frequencies associated with the anisotropic shielding interaction) for a sample in which a nucleus is described by a shielding ellipsoid $\boldsymbol{\sigma}$, and all orientations of $\boldsymbol{\sigma}$ are present for fixed B_0, will be

$$\Delta\omega = \gamma B_0(\sigma_{zz} - \sigma_{xx}).$$

Thus, a powdered sample of solid benzene, C_6H_6 (solid), in which all protons are chemically identical in that they are all aromatic protons on a single benzene ring, will have a proton NMR spectrum associated with the shielding interaction that is a powder average of the individual lines associated with the specific orientations for each benzene molecule. This spectrum has been found to be roughly 6 ppm wide and looks roughly as shown in Fig. 4e, Section IV. What happens now when the solid sample is converted to a liquid? The benzene molecules are free to rotate isotropically in solution. This means that the shielding ellipsoids characterizing the NMR lines associated with the shielding interaction are rotating isotropically. When the rotation frequency is faster than the spectral width characterized by the difference $(\sigma_{zz} - \sigma_{xx})$, then the observed shielding frequency is characterized by the isotropic average $\sigma_{\text{iso}} = (\sigma_{xx} + \sigma_{yy} + \sigma_{zz})/3$, and the observed NMR spectrum associated with the shielding interaction is a single line at angular frequency $\omega_{\text{iso}} = \gamma B_0(1 - \sigma_{\text{iso}})$.

In exactly the same manner that the anisotropic shielding is represented by an ellipsoid with three unequal axes, the anisotropy of the other three interactions may be so represented. A complete specification of an anisotropic interaction would include the lengths of the three axes of the interaction ellipsoid, and the three angles that orient this ellipsoid with respect to some coordinate system, such as the molecular framework in which the nucleus resides, or a fixed coordinate system within the laboratory. Therefore, six independent pieces of information completely specify an anisotropic interaction in general.

B. Dipolar Interaction

In addition to the shielding field, mentioned previously, a nucleus in a molecule will experience a magnetic field associated with the fact that there are other magnetic nuclei present. These nuclei act as little magnets, or magnetic dipoles, and provide an additional local field to the nucleus in question. This field represents a classical "through space" interaction, and has no relation to the electronic charge cloud present in molecules. The shape of this "dipolar field" due to a nuclear magnet is of the form of the pattern that iron filings take when spread around a bar magnet. Recall that this pattern varies both with direction, and distance from the magnet. Thus the effect of this field upon a neighboring nucleus will depend on where this neighbor is located within the dipolar field. The shift in NMR frequency due to this dipolar field will similarly depend on where the nucleus feeling the field is with respect to the nucleus producing the field. The observed shift in angular resonance frequency $\Delta\omega_{\text{dip}}$ of a nucleus in the presence of another nuclear magnetic dipole has spatial dependence

$$\Delta\omega_{\text{dip}} = \text{const} \cdot (1 - 3\cos^2\theta)/r^3,$$

where θ is the angle between the line connecting the interacting nuclei and the external magnetic field, and r is the internuclear distance. The constant is proportional to the magnitudes of the magnetic moments of the interacting nuclei. We thus see that the resonance frequency shift due to the dipolar interaction, if it could be measured without the interference of other interactions, provides a measure of nuclear geometries. It is notable that an interacting pair of dipolar nuclei have their resonance frequencies shifted by the inverse cube of their internuclear distance, so this shift is very sensitive to distance. Note that the above spatial dependence of the dipolar interaction between two nuclei does not contain the azimuthical angle ϕ. This statement translates into the fact that the ellipsoid describing the anisotropy of the dipolar interaction between two magnetic nuclei has two axes that are equal (i.e., there will be a plane in which all resonance frequencies due to the pairwise dipolar interaction will be the same). This result enforces a particular symmetry upon the NMR spectrum associated with a powdered sample of interacting dipolar pairs, which is illustrated in Section IV, Fig. 3a. It is a fact that the isotropic value of the dipolar interaction, which is $D_{\text{iso}} = (D_{xx} + D_{yy} + D_{zz})/3$ is zero. This fact will be subsequently important when the effects of motion on the dipolar interaction are considered.

A large number of interacting dipoles in a powdered sample would yield many resonance frequencies that would reflect the powder average of the angular distribution, $(1 - 3\cos^2\theta)$, and the sum of all pairs i, j, with internuclear distance r_{ij}, so the spectrum of nuclei in such a sample due to dipolar interactions alone would be a broad, featureless spread just reflecting average geometries and distances, as shown in Fig. 4c.

C. Electric Field Gradient; Quadrupolar Nuclei

The electric field gradient is simply the change in electric field with direction due to the local distribution of nuclear and electronic charges at a particular point in space in which the nucleus in question is located. For example, a sodium ion in a sodium chloride crystal would see an electric field and an electric field gradient associated with the presence of all neighboring Na^+ and Cl^- ions. In this special case, the electric field gradient is zero because the crystal symmetry is cubic. If this nucleus is magnetic, but has spherical nuclear charge symmetry (spin $\frac{1}{2}$, e.g., ^{13}C), then it is unaffected by a field gradient. If the nuclear charge symmetry is not spherical (spin greater than $\frac{1}{2}$, e.g., ^{27}Al with spin $\frac{5}{2}$), it can orient in an electric field gradient, which is to say that its nuclear energy levels that determine the NMR spectrum are sensitive to the field gradient. The spectrum associated with the central, $\frac{1}{2}$–$\frac{1}{2}$ transition of ^{27}Al in an electric field gradient that has axial symmetry for a sample of a powdered solid is shown in Fig. 3f, Section IV. Thus, the NMR spectrum of a quadrupolar nucleus associated with the presence of a nonzero electric field gradient is a measure of both local nuclear, *and* electron-cloud geometries. The ellipsoid characterizing the spatial symmetry of the electric field gradient is in general completely asymmetric (i.e., $E_{xx} \neq E_{yy} \neq E_{zz}$). While in general the isotropic value of the electric field gradient ellipsoid is not zero, to a first approximation it may be taken to be so. This fact will be important in considering the effects of motion on the NMR of quadrupolar nuclei.

D. Scalar Coupling

At this point we have seen that nuclear resonance frequencies can be sensitive to just the total electronic distribution (shielding), or just to the local distribution of magnetic nuclei about the nucleus in question (dipolar interaction), or to the total distribution of nuclei and electron charge (interaction of a quadrupolar nucleus in an electric field gradient). The fourth interaction to which all nuclei respond has a physical origin slightly different than any of the previous. It is a type of dipolar interaction, but transmitted from one nucleus to the other through the electronic charge distribution in a molecule. More specifically, it is transmitted by just that portion of the electronic cloud that touches both of the interacting nuclei. It is therefore a measure of a portion of the total electronic charge cloud in molecules. Because it is a dipolar interaction, its spatial dependence is exactly the same as for the classical dipolar interaction between two nuclei, with powder spectrum shown in Fig. 4a, Section IV. However, its magnitude, which depends on different physical factors, is quite different, and in general smaller than the classical dipolar interaction for nuclei in the first two rows of the periodic table.

To summarize the discussion, a nucleus that is not moving has an NMR spectrum that is characterized by a number of anisotropic interactions. These interactions may each be thought of as ellipsoids, with principal axes yielding NMR frequencies associated with the axis of quantization being along that axis. On isotropic rotation with a rotational frequency fast compared to the spectral width associated with the anistropy of any interaction in question, the observed spectrum is a single line at the frequency specified by the istropic value of the interaction. For example, the two major interactions of protons in solid benzene are shielding and dipolar. The shielding interaction has a spectral width of about 6 ppm, or of 1,800 Hz at a proton resonant frequency of 300 MHz, and the dipolar interaction has a spectral width of roughly 20 kHz. On melting, the benzene molecules in the liquid are isotropically rotating at frequencies much faster than 20 kHz, with the result that the observed proton NMR spectrum is associated with the isotropic values of the shielding and dipolar interactions. Because the isotropic value of the dipolar interaction is zero, the observed spectrum does not reflect the dipolar interaction at all. It is simply shifted from some reference by the isotropic value of the chemical shift only. The scalar coupling then acts to further split the lines in the observed spectrum, as will be illustrated later. If the nucleus in question is a quadrupolar nucleus (e.g., ^{23}Na, with spin $I = \frac{3}{2}$, then in addition to the effects of shielding and dipolar interactions of, for example, Na in $NaNO_3$, on the NMR frequency of Na, there will be an effect of the local electric field gradient. However, since the isotropic value of the electric field gradient is almost zero, sufficiently rapid isotropic rotation of the molecular environment about the sodium nucleus, such as would be experienced in an aqueous solution of sodium nitrate, will result in a sodium NMR spectrum that again reflects primarily the isotropic values of the shielding and scalar coupling.

III. NMR SPECTRA OF LIQUIDS

In Section II we have seen that although the nucleus in a molecule has an NMR spectrum that is a reflection of the entire molecular framework and the anistropy of this framework about the resonating nucleus, in the liquid state the resonance frequencies are simply a reflection of isotropic shielding and scalar coupling. This fact simplifies the observed spectra of liquids relative to those of solids. In addition, NMR has the capability of yielding both a quantitative and a qualitative analysis simultaneously. The reason for this remarkable fact is that it is

possible to uniformly excite all of the nuclei in a liquid sample in an NMR experiment. The result is that the total nuclear magnetization observed in an NMR experiment is proportional to the number of the nuclei in each chemical environment in all molecules under observation. Such is not the case for spectroscopies between the infrared and ultraviolet regions of radiation, which monitor rotational, vibrational, and electronic energy levels in molecules. In these spectroscopies, different chemical functionalities within a given molecule can have vastly different transition probabilities (i.e., extinction coefficients). This is the reason such spectroscopies, while diagnostic qualitative tools, are not convenient quantitative tools without extensive calibration procedures.

In addition, for NMR of liquid samples, again unlike visible absorption spectroscopies, the spectra of molecules are especially simple and easy to interpret. For example, the chemical shift NMR spectrum of ^{13}C in adamantane, $(CH)_4(CH_2)_6$, consists of two peaks roughly 10 ppm apart. Figure 2a indicates the ^{13}C NMR spectrum of adamantane (bottom), its carbon framework structure, and the time decay of the magnetization under a pulse NMR experiment (see Section IV) used to obtain the spectrum in the Fourier transform mode. We shall have more to say about pulse and Fourier transform NMR in a bit; for the moment, just concentrate on the spectrum and the structure. This structure has four methyne bridgehead carbons, H—C, and six methylene carbons, CH_2. A quantitative analysis of the areas of the two peaks shows that the areas are in the ratios 4:6, or 2:3, the same ratio as for the numbers of carbons with the chemical identification indicated.

The ^{13}C NMR spectrum shown for adamantane in Fig. 2a has been deliberately broadened to illustrate just the chemical shift spectrum in the absence of scalar coupling. Nevertheless, one sees that the two peaks are 10 ppm apart, with widths of roughly 5 ppm. This is quite a fantastic resolution, being able to see differences in absorptionfrequencies in the parts per million. With a magnet having a field of 14 T, carefully made uniform in order to limit the NMR linewidths to the minimum possible value

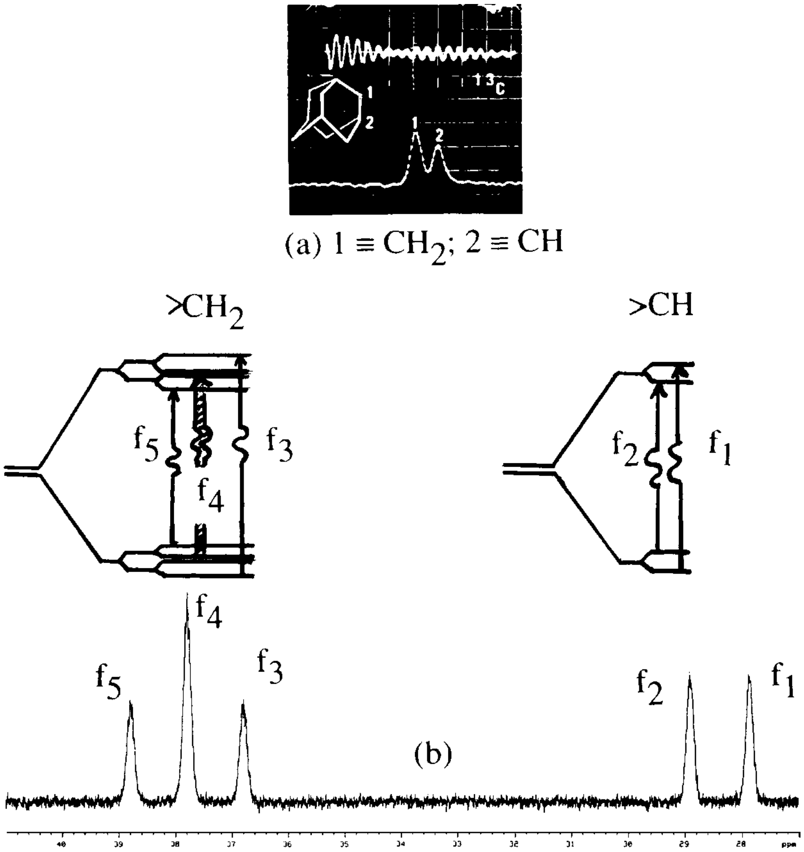


FIGURE 2 (a) The NMR of ^{13}C in adamantane. Top, time decay; center, adamantane; bottom, NMR spectrum. The 3 : 2 ratio of the areas of the peaks represents the quantitative ratio of methylene to bridgehead methyne carbons. (b) Spin–spin splitting at ^{13}C by ^{1}H in adamantane. Each proton acts as an extra field seen by the carbon, with direction along, and against the external field. H splits C-doublet. The two methylene protons split the methylene carbon signal into a triplet with intensity ratio 1:2:1. [High-resolution spectrum of ^{13}C in adamantane courtesy of Dr. Serge Lacelle.]

of the lifetime of the states involved in the transition, the attainable resolution has been 0.001 ppm, or a part per billion. This resolution was the origin of the analogy of being able to see two meter sticks half a meter apart on the moon. If a mixture of adamantane and methanol were to be the sample under investigation, with a resolution similar to that exhibited for adamantane alone, the liquid-state NMR of ^{13}C in this sample would reveal three peaks, one for each chemically different carbon in the sample, and the intensities and peak locations would lead one to correctly identify the sample as a mixture of the two chemicals, with relative amounts indicated by the peak intensities. Chemists, and more recently biochemists, have been utilizing this capability of liquid-state NMR now for almost thirty years. No chemical laboratory is complete without an NMR machine for structural studies. Almost every paper in the scientific literature reporting a new compound containing C and H that can be put in solution will have an NMR analysis as a routine part of the manuscript. As one further example of the use of NMR to elucidate the structure of molecules in solution, the proton NMR of phenyl epoxide Novolak resin is shown in Fig. 3. The interpretation of the spectrum is given in the legend of the figure.

Now we consider the effect of scalar coupling on the NMR spectrum of adamantane. We recall that the isotropic value of the scalar coupling is nonzero, so this value is an observable in the NMR spectrum of a nucleus in a molecular liquid. This interaction may be thought of as resulting from nuclei other than the nucleus under consideration giving rise to small magnetic fields transmitted via the bonding electron glue to the nucleus in question. These small magnetic fields act further to split the energies of the nucleus being considered, giving rise to further splitting of the spectrum that is quite diagnostic of the type and number of nuclei causing the splitting. For example, let the "nucleus in question" be the bridgehead methyne carbon, CH, in adamantane. Let the nucleus giving rise to the splitting of this carbon be the attached hydrogen,

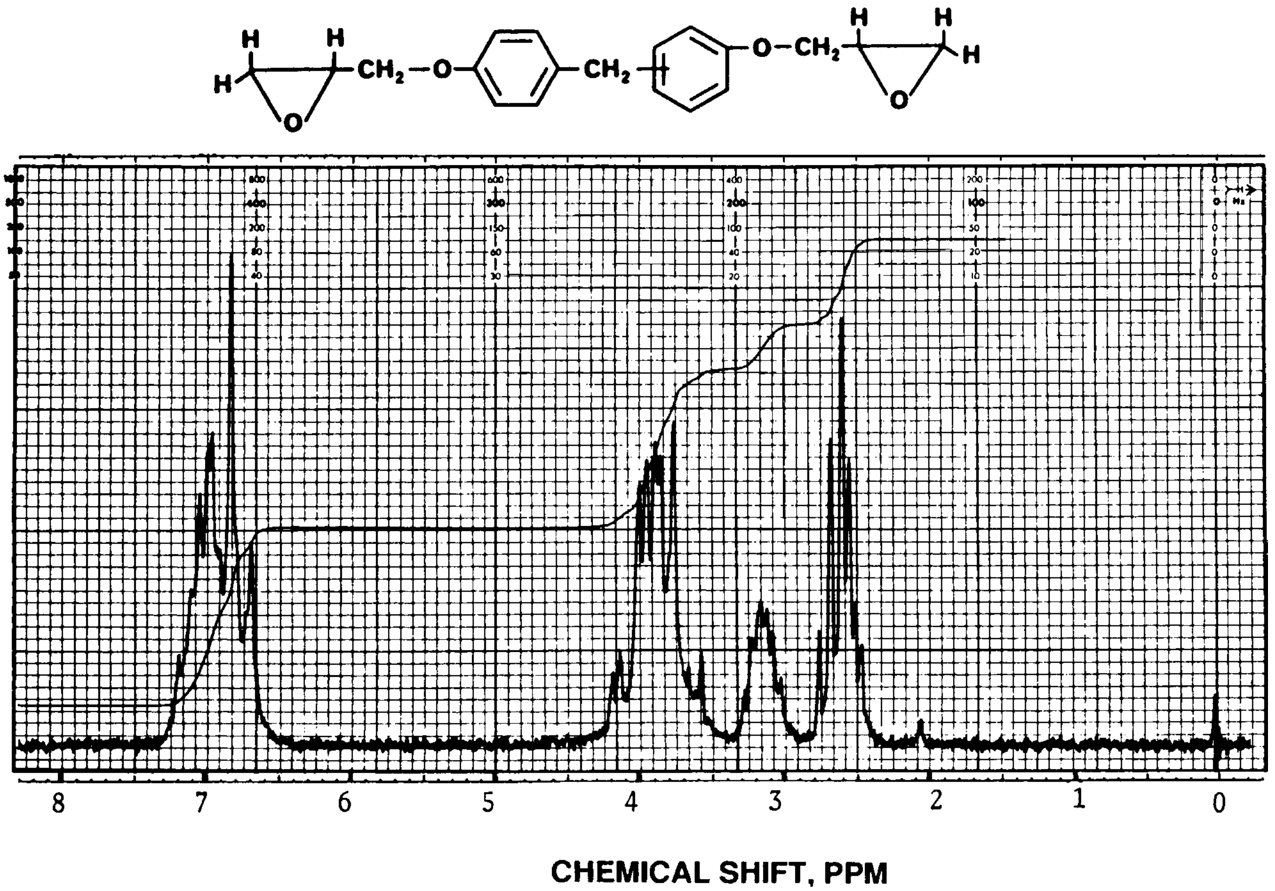


FIGURE 3 Proton NMR of phenyl epoxide Novolak resin. The monomer is shown above the spectrum. The line increasing in value running from right to left is the integrated value of the peaks in the spectrum, thus providing a quantitative analysis of the different types of hydrogen present in the polymer. The aromatic protons peaks at 6.69 and 6.83 ppm indicate that the rings are para substituted. The bridging methylene (CH_2) protons overlap with the epoxide band at 4.0 ppm.

with spin $\frac{1}{2}$, which means that the proton acts as a small magnet lining either with or against the external magnetic field. Then, as indicated in Fig. 2b, the methyne carbon resonance will be split into a doublet, with frequencies f_1 and f_2, centered about the original unsplit signal shown in Fig. 2a. Indicated at the right-hand side of Fig. 2b is the energy-level diagram of the methyne carbon, split by its lone proton, and the transitions, which are vertical lines with a wiggle in the center. These transitions are at frequencies f_1 and f_2, and are seen in the high-resolution spectrum of adamantane shown at the bottom of Fig. 2b.

At the left side of Fig. 2b are shown three peaks corresponding to the methylene carbon CH_2 in adamantane, split by its two attached protons. The manner in which these two protons split the carbon energy levels is indicated at the center of Fig. 2b. One proton supplies the same type of splitting as shown for the methyne carbon, and the second proton further splits that level as indicated. The transitions associated with this splitting are indicated by the "bathroom tile" diagram, at frequencies f_3, f_4, and f_5. Note that there are two equivalent transitions at f_4, so the intensity of the center line will be twice that of the two satellites. This experimental result is found in the high-resolution spectrum, shown on the lower left side of Fig. 2b.

Scalar coupling, or "spin–spin splitting" as it is commonly termed, is thus a diagnostic tool that adds considerably to the power of the NMR spectroscopist to diagnose details of molecular structure.

Structure determination of molecules in solution is by far the most common use of NMR at the present time. The amount of space devoted to this important application of the technique in this discussion is thus not commensurate with its present use, but perhaps not out of line in view of future applications in other fields such as materials science and medicine.

IV. NMR IN SOLIDS

In Section II, we discuss the effects of local arrangements of the electron cloud and other nuclei about a resonating nucleus in a molecule that are responsible for the effective magnetic field seen by this nucleus, and thus for its NMR absorption spectrum. Each interaction is discussed in turn, as if it were the only interaction present, the justification being that many timesthe NMR spectrum of a given nucleus appears to be the result of one or two interactions. A case in point is the NMR spectrum of a nucleus in a liquid, discussed in Section III. In this case, all information relating to anisotropy of the local environment effectively disappears, and the resulting spectrum is simply due to the isotropic portions of the chemical shift and of the scalar coupling. It has been previously noted that at least six pieces of information characterize each anisotropic interaction responsible for nuclear resonance frequencies. Thus, with four interactions and six pieces of information each, there are 24 parameters that could in principle contribute to each observed NMR spectrum. For a nucleus in a solid, such as ^{27}Al in an aluminosilicate catalyst, all interactions will be present and will contribute to the shape and width of the spectrum. The resultant spectrum will be quite broad and featureless compared to the relatively sharp NMR absorption lines observed for aluminum in a liquid solution. In other words, resolution, and thus information, has been lost. While there are in principle 24 pieces of information available, they are not separable from each other in this case, and in fact lead to a loss of information because they obscure each other. However, in many cases, there is one major contributor to the NMR spectrum of a nucleus in a solid. For example, the major contribution to the proton NMR of hydrogen in gypsum, $CaSO_4 \cdot 2H_2O$, is the dipolar interaction between two localized protons. The major contribution to the NMR of deuterium in perdeuterated polyethylene would be the interaction of the local electric field gradient with the quadrupole moment of the deuterium. Thus the physical information inherent in the spectrum due to single contribution (internuclear distances from dipolar interactions, etc.) may be extractable.

It is therefore useful to examine the powder spectra associated with each of the interactions in a solid. These spectra are shown in Fig. 3. These "powder patterns" have quite characteristic forms for particular symmetries of the interactions, and from these forms the types of structural information mentioned in the introduction can be derived. For example, the powder pattern associated with the dipolar interaction of two localized spin-$\frac{1}{2}$ systems, with gyromagnetic ratios γ_1 and γ_2, is shown in Fig. 4a. The splitting in frequency $\Delta\nu$/Hz between the two sharp horns of this spectrum is simply related to the internuclear vector r_{ij} by the formula

$$\Delta\nu = 1.5h\gamma_1\gamma_2\big/\left(2\pi r_{ij}^3\right).$$

Thus, as stated in Section II, the frequency splitting of this famous "Pake doublet powder pattern" is proportional to the inverse cube of the internuclear separation and is quite sensitive to this separation. For example, the two isolated protons at a distance of 0.248 nm (2.48 Å) in the trichloroacetic acid dimer, $(Cl_3CO_2H)_2$, in the solid state, yield a Pake doublet with "horns" 11,771 Hz apart, independent of the magnetic field. A change in the interproton distance by 0.01 Å results in a frequency change of 80 Hz in the separation of the horns, and easily detectable change by NMR and an almost impossible change to detect using X-ray diffraction. The separations of localized pairs

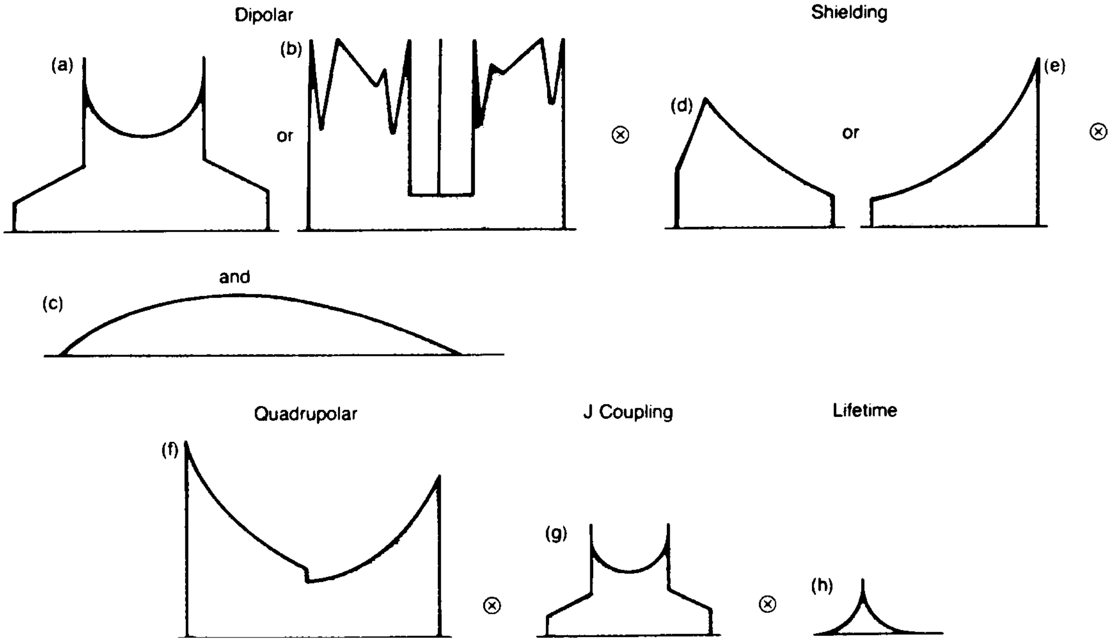


FIGURE 4 NMR powder patterns of nuclei. (a–c) Dipolar coupling. (d,e) Shielding anisotropy. (f) Central $\frac{1}{2}$–$\frac{1}{2}$ transition of a weak quadrupolar nucleus. (g) Scalar coupling anisotropy. (h) Lifetime broadening.

of protons in solids may be obtained from NMR much more easily than from X-ray diffraction. What does one do, however, when a given nucleus is exposed to a number of anisotropic interactions, all of which are roughly the same magnitude, such that the solid state spectrum is featureless and less informative than the information from just the chemical shift plus scalar coupling obtained in a measurement of NMR in the liquid state?

Fortunately, the introduction of transient techniques in NMR has led to the ability to remove, or attenuate some, and in favorable cases all, of the above intractions in a selective manner, such that all but the interaction desired to be seen remains. These techniques are outside the limitations of the present writing, but may be viewed in detail in some of the references supplied. An example of the removal of broadening due to proton–proton dipolar interactions, without removal of shielding anisotropy, in a powdered solid of high-density linear polyethylene (in which there is only one chemical species of proton, the methylene proton), is given in Fig. 5. Here, the powder spectrum without (a) and with (b) transient techniques used to separate proton–proton dipolar coupling from shielding anisotropy are shown. In part (a) of Fig. 4 is shown the powder spectrum of protons in polyethylene under a standard experiment in which the spectrum is a result of both shielding anisotropy and dipolar coupling. In this case, the dipolar coupling dominates the spectrum, and the line width is roughly 20 kHz. Also at the top the narrow line is indicated resulting from manipulating the dipolar interaction so as to remove it from the observed result. This is the quite narrow line inside the broad line. At the bottom is an enlarged version of the narrow line, where the abscissa has been changed from kHz to ppm. In the spectrum shown, the resonant frequency was 56.6 MHz, so one ppm is 56.6 Hz. The full width of half height (fwhh) is about 3 ppm, corresponding to about 170 Hz. Thus a spectrum 20,000-Hz wide has been reduced to a spectrum about 200-Hz wide by manipulating the nuclear spins in a manner now becoming available via transient techniques in NMR.

In a similar manner, each of the other broadening interactions can be manipulated to remove or scale their values such that the resulting spectrum reveals features associated with predominantly one interaction. This manipulation of spin and real-space operators is a small branch of science in itself, and the subject of a number of texts referenced at the end of this article. An interesting feature of these manipulations is the physical rotation of the samples at speeds about that of dentists drills: about 180,000 rpm. A favorite angle of orientation of the rotation axis with respect to the static field is the so called "magic angle," which is that along the diagonal of a cube with respect to a cube edge: 54.77°. If a sample is rotated at this angle in the static field at speeds large compared to the shielding anisotropy in cycles per second, the resulting broad powder spectrum associated with the shielding anisotropy is reduced to its isotropic value (i.e., to a sharp line comparable in width to spectra observed in liquids). Thus, for a system

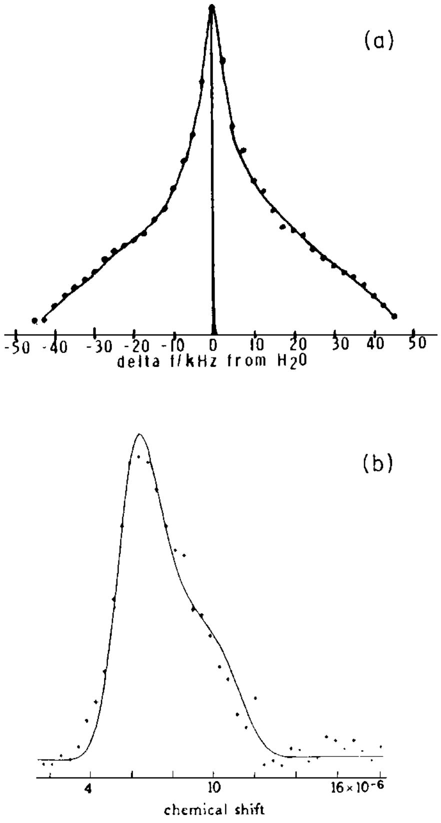


FIGURE 5 (a) Proton spectrum of polythylene taken in standard manner using multiple pulse decoupling to remove proton–proton dipolar interactions leaving just the shielding anisotropy. (b) An expanded view of (a). Note that the scale of (b) is in ppm (1 ppm = 56 Hz), and that of (a) is in kHz.

in which dipolar broadening and shielding anisotropy are the only two broadening interactions present, a combined experiment in which dipolar broadening is removed via exposure of the nuclei to some type of rf irradiation while the sample is rotated at a frequency fast compared to the shielding anisotropy will result in a liquid-like spectrum.

V. PULSE AND FOURIER TRANSFORM NMR

Implicit in some of the previous discussion is the fact that in modern instruments an NMR spectrum is obtained by exciting the system under study by a short pulse of radiofrequency power of duration t_p observing the time decay (see the decay in Fig. 2, which has been transformed to obtain, the spectrum of ^{13}C in adamantane) and performing a manipulation to obtain the spectrum from the time decay. The process of converting the time decay to the frequency spectrum is performed by a Fourier transform. The signal in the time domain, $S(t)$, is related to the signal in the frequency domain, $S(\omega)$, by the formula

$$S(\omega) = \left(\frac{1}{2}\pi\right) \int_0^\infty dt\, S(t) \exp[-i\omega t].$$

A tremendous advantage of pulse and Fourier transform NMR is the fact that the spectral width scanned in a single pulse of pulse width t_p is roughly $\frac{1}{2}t_p$. For example, a pulse a microsecond long has a spectral width of 500 KHz, meaning that all nuclei with resonant frequencies 250 KHz on each side of the central carrier frequency will be equally excited, and will experience resonant transitions. This fact has potent implications in obtaining spectra by signal averaging because an entire spectrum can be scanned in very few seconds for many nuclei. It is now possible, using transient techniques, to obtain spectra of nuclei simply unavailable in the past, and at the time of the present writing, almost every nucleus in the periodic table is available as a spin label for NMR studies. Figure 6 graphically illustrates the relation between the decay of the magnetization in the time domain (top) and the in and out of phase Fourier transforms leading to the absorption and dispersion spectra, respectively. A person walking into a laboratory in which a pulse NMR spectrometer is operating today is likely to see both of these signals on the video screen of the instrument; one being the signal under accumulation in the time domain, and the other the time-averaged signal which has been Fourier transformed to reveal the absorption spectrum as it is being accumulated (see Fig. 2a).

A special feature of NMR is that the quantum mechanics are frequently so simple that many of the problems have solutions that can be obtained in closed form. The solutions of these problems may also be visualized classically. For example, a pulse NMR experiment may be visualized as the result of creating a vector (the net nuclear magnetization after allowing a sample to come to

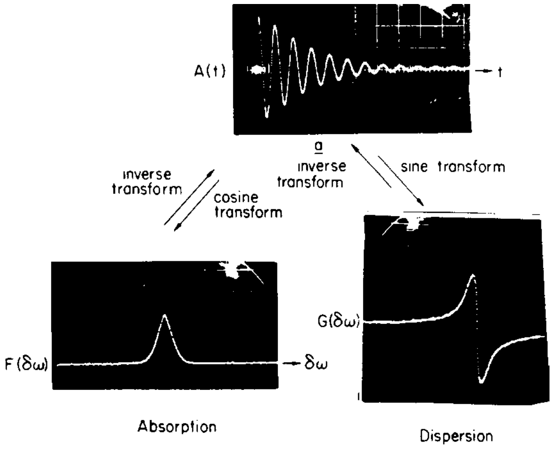


FIGURE 6 Relation between the time decay (top) in a transient magnetic resonance experiment, and the absorption (left, bottom) and dispersion (right, bottom) of the NMR signal.

equilibrium in a static magnetic field), then creating for a time t_p another field (the rf field), perpendicular to the static field. As indicated in the introduction, the basic response of a nucleus in a magnetic field is to precess about the field with a precession frequency $\omega = \gamma B$. Therefore, during a pulse with spectral width $\Delta\nu = \frac{1}{2}t_p$, all nuclei within this spectral width may be thought of as simply precessing about the B_1 magnetic field of the pulse with angular precession frequency $\omega_1 = \gamma B_1$. If the pulse is left on for a time t_p, then the precession angle Θ_p is given by

$$\Theta_p = \gamma B_1 t_p = \omega_1 t_p/\text{rad}.$$

If $t_p\omega_1$ set to $\pi/2$ radians, the nuclear magnetization will precess to a position perpendicular to its original orientation. At this point in time, it is then free to process around the static field B_0. In accord with classical magnetism, a rotating magnet creates a voltage in a coil arranged with its axis perpendicular to the axis of rotation of the magnet. This oscillating voltage is the nuclear induction signal that is observed as the time decay and in turn is transformed into the spectrum. A classical picture of the process just described is given in Fig. 7. At the top, the pulse field rotates the magnetization to the transverse plane. The experimenter views this magnetization by gazing at a fixed axis in this plane (this process is known as "phase detecting"). The moment will in general process about this axis (center, Fig. 7), giving rise to an oscillating signal detected by the experimenter (bottom, Fig. 7). This oscillation will in general be damped, with a time constant T_2^* such that the envelope of the oscillation is of the form $\exp[-t/T_2^*]$. The term T_2^* is called the *transverse, or spin–spin, relaxation time*. Its value offers an insight into motions of the sample in the zero frequency and $2\omega_0$ frequency range. The time constant characterizing the return of the ensemble of nuclear spins back to the direction of the static field is called the spin-lattice, or longitudinal relaxation time, T_1. Its value gives information about motion in the frequency range of the precession frequency of the spins in B_0, which is $\omega_0 = \gamma B_0$.

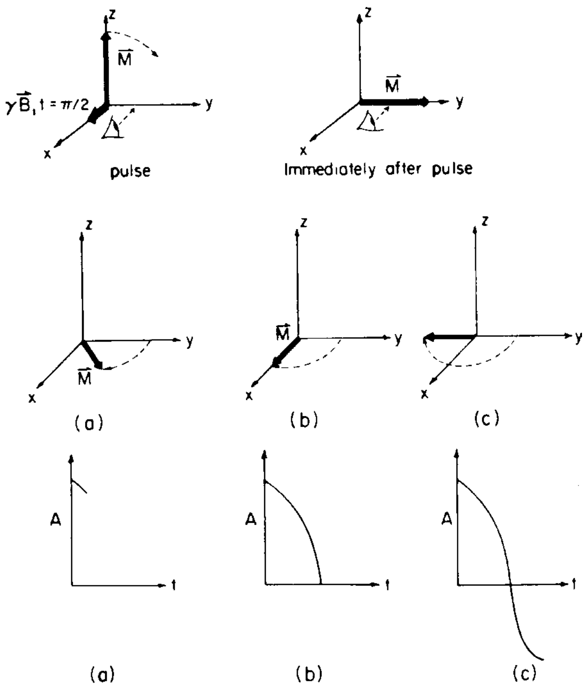


FIGURE 7 Classical picture of a pulse NMR experiment. Relation between precessing moment (top and center) and the observed transverse component of the magnetization as a function of time (bottom).

Pulse experiments can be performed that characterize other time constants, the description of which is beyond the scope of the present treatment.

In the previous discussion, we have concentrated on "one-dimensional" data acquisition; intensity versus frequency. There are multidimensional techniques available, which we now introduce.

VI. TWO-DIMENSIONAL NMR

In a one-dimensional NMR experiment, data are taken as a function of a single time parameter, and the relation between these data and the frequency spectrum is the previously discussed Fourier transform relation. Over the past few years, a number of experiments have been developed in which the time intervals in the NMR experiments are divided into regions, a region t_1, followed by another region, t_2. The time domain signal, then, is a function of both of these times; $S(t) \equiv S(t_1, t_2)$. An immediate result of this statement is that the frequency domain signal, $S(\omega_1, \omega_2)$, now becomes a three-dimensional contour plot, as shown in Fig. 8.

Figure 8 is a two-dimensional plot in which chemical shifts of the three different carbons in *n*-hexane, CH_3—CH_2—CH_2—CH_2—CH_2—CH_3, are plotted on the "ω_2" axis (going into the plane of the paper), and the chemical shifts-plus-spin–spin couplings are plotted on the "ω_1" axis (parallel to the plane of the paper). The "ω_1" plot is what one would obtain in a 1-D NMR experiment in which both chemical shifts and scalar (J) couplings are simultaneously present. The "ω_2" plot is what one would obtain in a 1-D experiment in which the scalar couplings of the protons to the carbons are averaged to zero by what is called "decoupling," accomplished by irradiating the proton frequencies while the carbon signal is observed. Clearly, there is less information on the ω_1 and the ω_2 axes than in the 2-D plot shown in the plane, where it is obvious which chemically shifted carbons are attached

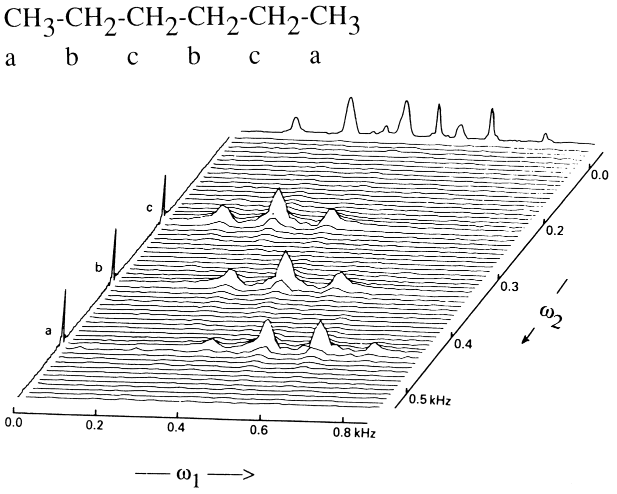


FIGURE 8 Two-dimensional plot of ^{13}C NMR. The peaks along the ω_1 axis (at the top of the figure) are the ^{13}C resonances taken without decoupling the protons from the carbons. Overlapping peaks make spin–spin coupling multiplicities hard to determine. The peaks along the ω_2 axis are the ^{13}C resonances obtained when the protons are decoupled from the carbons, thus eliminating spin–spin coupling multiplicities. The peaks in the 2-D area clearly show which chemically shifted carbons have two or three protons attached. A triplet represents two attached protons, a quartet three.

to how many protons. For example, the chemical shifted carbon labeled "a" at a frequency of roughly 0.45 kHz on the ω_1 axis is clearly split into a quartet in the 2-D plot. This uniquely identifies this carbon as being attached to three chemically equivalent protons, or as a CH_3 group. The chemically shifted carbons labeled "b" and "c" on the ω_1 axis are J-split into triplets, which clearly indicates, as with the previously discussed case of adamantane, that they are methylene carbons.

This is one of the simplest 2-D experiments possible, and there is an entire catalogue of such experiments which can be used, for example, to trace connectivities of carbon–carbon bonds through a chemical structure. The latter is especially useful in complicated structures, such as are found in molecules such as steroids and proteins, of great interest to biochemistry and the medical profession.

VII. MULTIPLE QUANTUM NMR

A. Coupled Spins $\frac{1}{2}$

In the discussion thus far, we have been dealing with what is termed "single quantum" NMR. This term is used because the quatum mechanical states that result in the observed signal differ by unity, that is, by a "single" quantum of angular momentum, in the quantum number that relates to the nuclear angular momentum, and thus to the magnetic moment. This is to say that the observed signal in the time domain, such as shown in Fig. 2, is the signal associated with ^{13}C in adamantane resulting from a single pulse excitation (Fig. 7). The time-dependent state leading to this time-dependent signal is a superposition of two states differing by unity in the quantum number describing the nuclear angular momentum. The observed signal, of course, is the average over all the spins in the sample, and thus may be said to be a *coherent magnetization of all spins in the ensemble of molecules being observed*. Another way of speaking of this signal is that it represents a *single quantum coherence*.

It is possible for systems of nuclear spins to have states that differ by greater than unity in that quantum number for angular momentum. There are at least two ways in which this may come about

The first way is two or more spins $\frac{1}{2}$ may couple together to form a spin greater than $\frac{1}{2}$. For example, two spins $\frac{1}{2}$ may couple to form a system with spin unity. In this case, there are $2S + 1 = 3$ states describing this coupled system, and there can be two states that differ by two in angular momentum. If an excitation is provided which produces a time-dependent state that is a superposition of the two states differing by two in angular momentum, then with that superposition being coherent over the ensemble of spins being observed, we speak of a *double quantum* coherence. Extending the argument to many coupled spins $\frac{1}{2}$ leads to the idea that multiple quantum coherences may be produced by suitable excitation. In fact, the highest order of coherence being observed is one means of counting the number of spins that are coupled. As one example of the use of such an idea on which the author has worked, consider the problem of determining how many protons there in a fragment of ethylene adsorbed on a metal catalyst during any part of the processes involving adsorption and reaction. These are disordered systems, with no coherence length amenable to analyses by X-rays. One possibility is to count the maximum number of coupled protons resulting from such an adsorption and subsequent reaction. In this case, it was found that at a particular time in the reaction sequence, there were six coupled protons in the C_nH_m fragment bound to the surface, which indicated that the fragment could be (Ru)—CH_2—CH═CH—CH_2—(Ru), where the Ru in parenthesis stands for the metal in the catalyst particle.

The multiple quantum *spectrum* associated with this six-quantum coherence is shown in Fig. 9. The intensities of the peaks associated with higher orders of coherence drop off roughly as a Gaussian curve, and the highest order of coherence developed depends upon the time allowed for this development. In the present example, only five peaks were developed no matter how long time was allowed for higher coherences to develop, indicating that with the

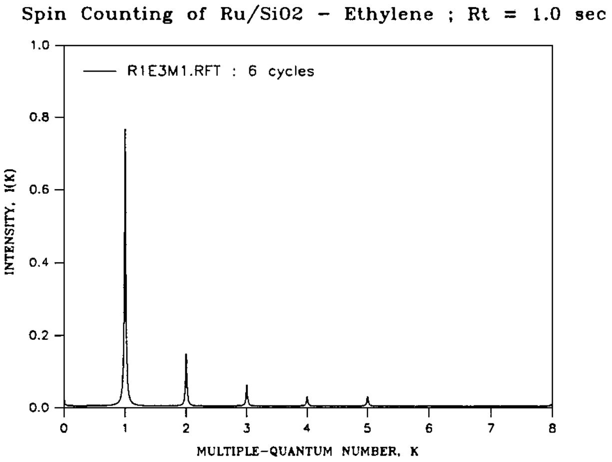


FIGURE 9 Multiple quantum spectrum of coupled protons in a C_nH_m fragment resulting from the adsorption and reaction of ethylene on a supported ruthenium surface. The plot is of intensity versus order of coherence. The highest order, *k*, observed, is 5, meaning that in this case there are six coupled protons in the sample. The inference is that the fragment being observed could be (Ru)–CH_2–CH=CH–CH_2–(Ru), where the (Ru) indicates the metal surface.

selection rules of the experiment, there are six coupled protons. This figure is *not* a plot of intensity versus *frequency*. Rather, it is a plot of intensity of each k quantum coherence versus the order of the coherence, k. The method of obtaining this information from the data, which in pulsed experiments are always represented by a plot of intensity versus time, is beyond this article. It must be emphasized, however, that the observable in NMR experiments is *always* single quantum coherence. The implication is that to obtain the information in which the multiple quantum coherence is contained in the observable, it is necessary to convert from multiple to single quantum coherence in a two-dimensional NMR experiment, one example of which was given in Section VI.

A rather lovely aspect of this experiment is that there is a sense in which time is reversed, in that the experiment starts with the system being in a state of Zeeman order (all spins aligned along the external field), is caused to develop multiple quantum coherence, and then caused to reverse its evolution back into a state of Zeeman order again. NMR is a wonderful tool for the study of time-dependent quantum mechanics.

B. Quadrupolar Nuclei

A second case in which it is possible to have quantum states that differ by greater than unity in angular momentum quantum number is that of nuclei with spin greater than $\frac{1}{2}$. Almost 70% of the nuclei in the periodic table have this property. Some are quite common and therefore important to chemists, materials scientists, and others who may have an interest in analyzing the properties of matter on a microscopic scale. Examples include ^{7}Li, with spin $\frac{3}{2}$, Li being an important part of the current-carrying species in Li batteries; ^{27}Al, with spin $\frac{5}{2}$, Al being a critical part of catalysts used for oil refining; and ^{11}B, with spin $\frac{3}{2}$, boron being an important species used in production of new glasses for use in fiber optics.

Each nucleus with spin S can have $2S+1$ quantum states that describe its angular momentum. So, for example ^{27}Al has four states describing its angular momentum, and it is possible to produce three-quantum coherence in ensembles of such nuclei, no coupling being present.

The importance of multiple quantum coherence in such systems lies in the fact that use is made of such coherences to produce high-resolution NMR of these species in *solid*. Until these techniques were developed in the late 1990s, high-resolution NMR of quadrupolar nuclei in solids, with the resultant power of the chemical shift as a fingerprint, was unachievable in a manner that was technically simple enough to be used by the average NMR spectroscopist. This “holy grail” of solid state NMR has now been found, and spectra of species such as ^{87}Rb in the solid state have been obtained with a resolution of a single part per million, as illustrated in Fig. 10. The conditions under which the bottom spectrum of Fig. 10 was achieved involved spinning the sample at 1.2 million rpm, at the so-called magic angle to the static field, and, simultaneously creating, and detecting multiple quantum coherence. The experiments have been labeled MQMAS, for *M*ultiple *Q*uantum *M*agic *A*ngle *S*pinning. This is, in the author’s opinion, a stunning achievement with implications to be developed far into the future.

VIII. MEDICAL IMAGING

The precession frequency of nuclear spins in a magnetic field has been alluded to now a number of times; it is $\omega_{\mathrm{eff}} = \gamma B_{\mathrm{eff}}$. A physical result of this statement is that if B_{eff}, the effective magnetic field at the nucleus, can be made to vary in space, then the resonant frequency will vary in space. This fact is the basis of imaging of nuclei in the human body, and for that matter, in any region of matter. In an imaging experiment, coils of wire are placed about a subject in a magnetic field, and pulsed current is run through these coils to produce gradient magnetic fields that vary with space and time. The nuclear resonance frequencies of nuclei in the subject in turn vary with space and time. Since different tissues (e.g., muscle compared to bone) have different concentrations of nuclei such as protons, as well as differing relaxation times T_1 and T_2, for these nuclei it is possible to convert a

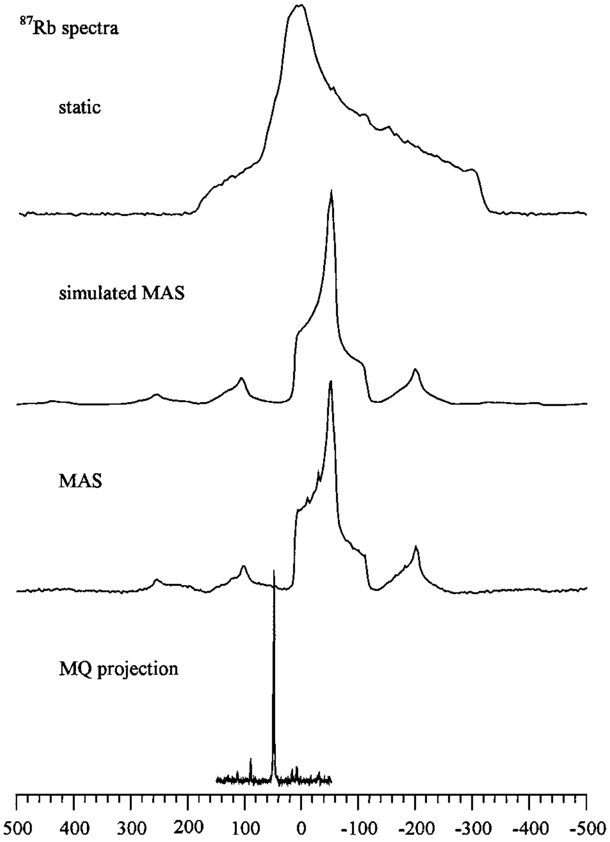


FIGURE 10 NMR spectra of the spin $\frac{3}{2}$ quadrupolar nucleus ^{87}Rb, taken under (top) static, (middle two) magic angle spinning, and (bottom) conditions of magic angle spinning, and multiple quantum coherence (MQMAS).

three-dimensional net of intensities of nuclear resonance frequencies in various tissues into images representing the tissues themselves. The physician can then call any two-dimensional slice of this information such that sections of the human body can be viewed, appropriately colored, on a television monitor. The information may also be converted to colored photographs that rival photographs of actual organs in their appearance, and are in general of higher resolution than is achievable from X-ray films. The general scheme of a patient in the machine is shown in Fig. 11. One such slice of information taken from a sagittal scan through the eye is shown in Fig. 12.

As a further example of the ability of NMR to form an image of a section through the human body without the use of damaging ionizing radiation, Fig. 13a and b show whole-body scans. Figure 13a is a section through the upper chest region perpendicular to the spinal cord. The patient is prone, and the spinal column is seen at the bottom center of the scan. The two upper arms, including muscle, fat, and bone, are seen on either side of the torso to the right and left. Figure 13b is another section of the same individual, but this time taken parallel to the spinal cord,

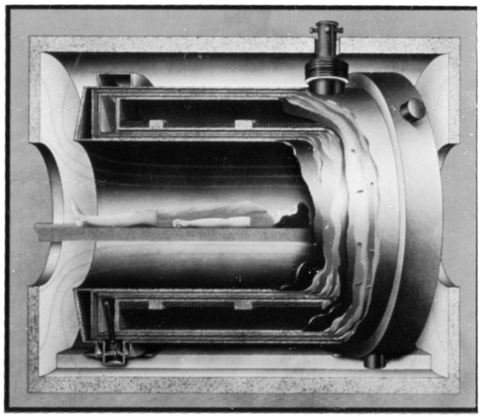


FIGURE 11 Arrangement of a magnet and patient for a whole body imaging scan. [Courtesy of Wang NMR, Inc.]

seen as the vertical column in the center. Clearly seen are the lungs, portions of the ribs, and a detailed picture of the spinal column.

Imaging is in its infancy, and given the fact that radiofrequency radiation is nonionizing, it is likely that such a technique will be widely used in lieu of X radiation for specific applications in which sensitivity of the body to X-rays is a problem. Also, since NMR is nucleus specific, whereas X-ray scans see only dense versus nondense matter, the diagnostic potential of NMR imaging is quite promising.

For example, the use of ^{31}P as an NMR tag to detect concentrations of creatine phosphorus in the heart of a patient after a coronary infarction may be used to diagnose the extent of the damage to the heart muscle.

If one examines the statement just made carefully, it may be seen that the entire discussion of the utility of NMR to probe materials lies in the fact that nuclei have a number

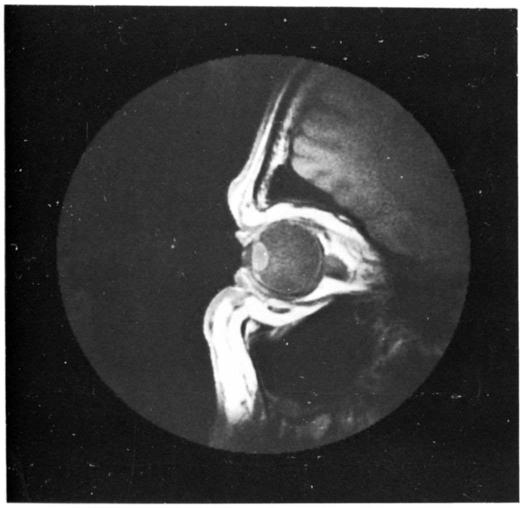


FIGURE 12 Sagittal scan through the eye region of a human. [Courtesy of Dr. John Schenck, General Electric Company.]

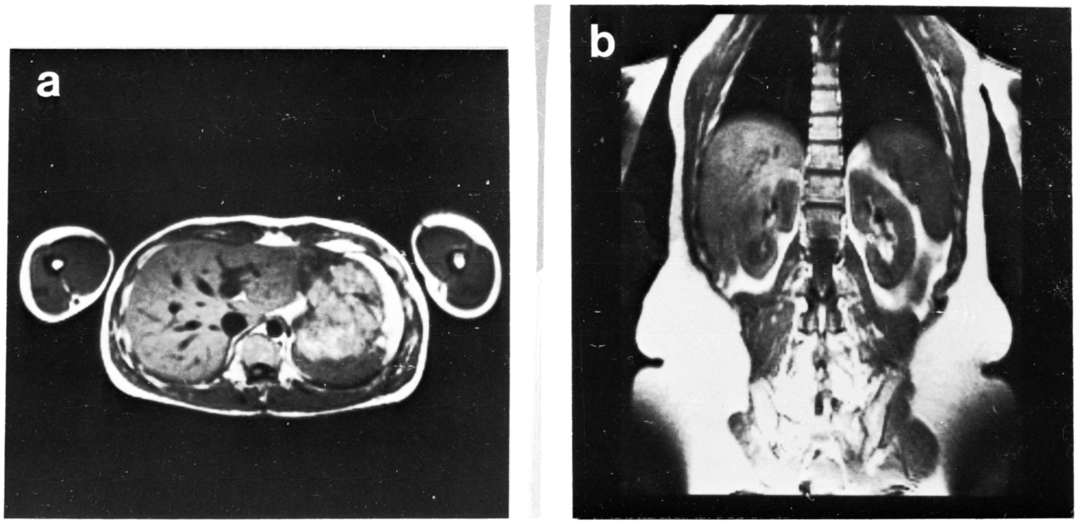


FIGURE 13 (a) Scan through upper trunk perpendicular to spine. (b) Scan through upper trunk parallel to the spinal cord. [Courtesy of Dr. John Schenck, General Electric Company.]

of characteristic fingerprints that can be used to probe their environments. We have talked about fingerprints associated with shielding, dipolar interactions, quadrupolar interactions, and their reflections as seen in the various relaxation times available to the nucleus. The field of imaging is just beginning to use these fingerprints for enhanced resolution. For example, there now exist "T_1 images", and "T_2 images" that use the fact that transverse and longitudinal relaxation times of nuclei in a given tissue are characteristic of that tissue. The full range of interactions of nuclear behavior has yet to be exploited for imaging.

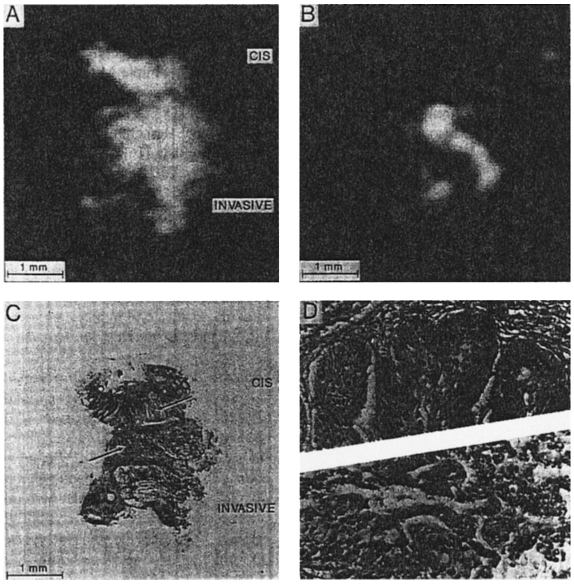


FIGURE 14 Chemical shift imaging allows both the location and pathology of a cervical lesion to be obtained simultaneously. Water based (a) and Lipid based (b) images of two human cervical biopsies are shown. The lower biopsy is an invasive carcinoma and the upper biopsy is a CIS/CN3. Distinction between the two pathologies is apparent, with bright areas detected in the lipid image of invasive carcinoma but not CIS/CIN3. (c) A 5 μm histological cut through the two biopsies parallel to, and at the center of, the imaged slice. (d) Magnification ($\times$10) of the tissue area indicated by the arrows in (c). The top and bottom panels show the histology of CIS and the invasive carcinoma, respectively.

Two recent developments which illustrate the burgeoning power of NMR imaging (MRI) in medical, biological, and materials science are (1) noninvasive diagnosis of cancer by so-called chemical-shift imaging; (2) imaging of live silk-butterfly pupae growing inside the cocoon; and (3) imaging with the imaged body (live and human, or inanimate, and a rubber band) *outside* of the magnetic system, via the so-called NMR "Mouse." Examples of each type of image is shown in Fig. 14.

ACKNOWLEDGMENT

The author's research is supported by The Energy and Minerals Resources Research Institute, operated for the U.S. Department of Energy by Iowa State University under Contract Number W-7405-Eng.82. Dr. C. R. Dybowski and Shelly Ironside helped to provide a modern update of the current article.

SEE ALSO THE FOLLOWING ARTICLES

ANALYTICAL CHEMISTRY • MACROMOLECULES, STRUCTURE • MAGNETIC RESONANCE IN MEDICINE • ORGANIC CHEMISTRY, COMPOUND DETECTION • STEREOCHEMISTRY

BIBLIOGRAPHY

Becker, E. D. (1980). "High Resolution NMR—Theory and Applications," 2nd ed. Academic Press, New York.

Ernst, R. R., Bodenhausen, G., and Waukon, A. (1987). "Principles of Nuclear Magnetic Resonance in One and Two Dimensions," Oxford Univ. Press, Oxford.

Farrar, T. C., and Becker, E. D. (1971). "Pulse and Fourier Transform NMR," Academic Press, New York.

Fraissard, J. P., and Resing, H. A. (eds.). (1980). Magnetic resonance in colloid and interface science, *NATO Adv. Study Inst. Ser, Ser C: Mathematical and Physical Science* **61**.

Fukushima, F., and Roeder, S. B. W. (1981). "Experimental Pulse NMR: A Nuts and Bolts Approach," Addison-Wesley, Reading, MA.

Fyfe, C. A. (1983). "Solid State NMR for Chemists," C. F. C. Press, Guelph, Ontario.

Gerstein, B. C., and Dybowski, C. R. (1985). "Transient Techniques in NMR of Solids: Introduction to Theory and Practice," Academic Press, Orlando.

Harris, R. K., and Mann, B. E. (1976). "NMR and the Periodic Table," Academic Press, New York.

Hwang, S. J., King, T. S., and Gerstein, B. C. (1991). "Probing intermediates in the reaction of ethylene over supported Ru: Utility and limitations of the use of multiple quantum coherence," *Catalysis Letters* **8,** 367–373.

Macomber, R. S. (1988). "NMR Spectroscopy: Essential Theory and Practice," Harcourt Brace Jovanovich College Outline Series, Harcourt Brace Jovanovich, San Diego.

Mehring, M. (1981). "High Resolution NMR of Solids," Springer, Heidelberg.

Mountford, C., Doran, S., Lean, C. L., and Russell, P. (1980). "Cancer pathology in the year 2000," *Biophys. Chem.* **68,** 127–135.

Petrakis, L., and Fraissard, J. P. (eds.). (1984). "Magnetic resonance: Introduction, advanced topics, and applications to fossil energy," *NATO Adv. Study Inst. Ser., Ser. C, Mathematical and Physical Sciences* **24**.

Roeder, S. B. W. (1981). "Experimental Pulse NMR: A Nuts and Bolts Approach," Addison-Wesley, Reading, MA.

Slichter, C. P. (1978). "Principles of Magnetic Resonance," 2nd ed., Springer, Heidelberg.

Organic Chemistry, Compound Detection

Raphael Ikan
Bernard Crammer
Hebrew University of Jerusalem

I. Introduction
II. Chromatographic Methods
III. Flash Chromatography
IV. Droplet Countercurrent Chromatography
V. High-Pressure Liquid Chromatography
VI. Gas Chromatography
VII. Supercritical Fluid Chromatography
VIII. Thin-Layer Chromatography
IX. Field-Flow Fractionation (FFF)
X. Ion Chromatography (IC)
XI. Capillary Electrochromatography (CEC)
XII. Spectroscopic Methods
XIII. X-Ray Crystallography
XIV. Chemical Analysis
XV. Future Prospects

GLOSSARY

Capillary electrochromatography (CEC) Technique wherein a packed bed capillary with smaller packing particles than in HPLC systems is used. It uses electroosomotic flow (EOF) rather than high pressure to force the mobile phase through the capillary.

Chromatography Method of separating two or more substances by distribution between two phases, one fixed (the stationary phase) and the other moving (the mobile phase).

Chromatography of isomers Chromatographic separation of geometric (cis-trans) and optical (*R* and *S*) enantiomers on adsorbents that have been impregnated with compounds having the ability to complex preferentially or interact with specific functional groups causing the required separation.

Coupled chromatographic and spectroscopic techniques Method involving a chromatographic system such as GLC or HPLC connected to a mass spectrometer or FTIR spectrometer in order to analyze individual components from a mixture of organic compounds in minute (submilligram) quantities.

Deuterium exchange Replacement of hydrogen atoms by deuterium atoms usually by means of active compounds containing deuterium such as D_2O and $NaBD_4$. The percentage of deuterium exchange may be determined by spectroscopic analysis.

Diffuse reflectance IR spectroscopy (DRIFTS) Penetration of infrared radiation onto a solid sample before the radiation is scattered.

Droplet countercurrent chromatography (DCCC) Separation technique based on liquid–liquid partition chromatography.

Flash chromatography Also known as rapid column chromatography.

Fourier transform Technique in which a short powerful radio-frequency pulse (microseconds) excites either all the ^{1}H nuclei or ^{13}C nuclei simultaneously. Each nucleus shows a free induction decay (FID) which is an exponentially decaying sine wave with a frequency equal to the difference between the applied frequency and the resonance frequency for that nucleus.

Gas–liquid chromatography (GLC) Technique in which the organic sample is carried through a column by a carrier gas (mobile phase) and the separation of the organic compounds occurs in the stationary phase (the column packing). The compounds are estimated by means of a detector.

Gas–solid chromatography Technique in which the moving phase is a mixture of gases and the stationary phase is a solid phase. The carrier gas such as nitrogen or helium replaces the solvent in column chromatography. The solid may be finely powdered Celite or kieselguhr. The technique is suitable for organic substances that are volatile without decomposition up to about 300°C.

Gel permeation chromatography (GPC) Technique that separates substances according to their molecular size and shape. Three classes of stationary phases are used: aerogels (porous glass), xerogels (crosslinked dextran), and xerogel-aerogels (polystyrene).

^{1}H NMR "polarimetry" Technique using chiral shift reagents in NMR spectroscopy for the determination of enantiomeric purity of natural semiochemicals.

Heteronuclear shift correlation (HETCOR) 2D-NMR technique where a two-dimensional map is generated in which one frequency axis corresponds to ^{13}C chemical shift or other nonproton nuclei, and the other frequency axis corresponds to the ^{1}H chemical shift.

High-performance thin-layer chromatography (HPTLC) Technique enabling the separation of very complex mixtures of organic compounds. The plates are prepared from optimized thin adsorbent layers. HPTLC offers greater separation efficiency through smaller plate heights than the conventional TLC plates, shorter analysis time, and detection limits in the nanogram and picogram range.

High-pressure liquid chromatography (HPLC) Technique consisting of a stationary phase (a solid surface, a liquid, an ion-exchange resin, or a porous polymer), held in a glass or metal column with the liquid mobile phase being forced through under pressure.

Homonuclear shift correlation (HOMCOR) 2D-NMR technique where a two-dimensional map is generated in which both of the frequency axes corresponds to homonuclear shifts in a single spectrum. Also known as COSY.

Infrared spectroscopy Technique in which many functional groups and atoms are characterized by their vibrations and deformations in the 4000–200 cm^{-1} range.

Ion-exchange chromatography (IEC) Technique in which the stationary phase consists of a rigid matrix (polymer), the surface of which carries a net positive (cationic) or negative (anionic) charge to give an ion exchange site R^{+} or Y^{-}, respectively, which will attract and hold the counterions. Ion exchangers are divided into *anion* and *cation* exchangers.

Lab on-a-chip This technique uses a planar device on which a number of chemical processes are being performed in order to go from reactants to products or from a sample to analysis.

Liquid–solid (absorption) chromatography (LSC) Separation of compounds (from a mixture) by a liquid mobile phase and a solid stationary phase which reversibly absorbs the solute molecules.

Mass spectrometry Technique in which a vaporized sample of a substance is bombarded with a beam of electrons, and the relative abundance of the resulting positively charged molecular fragments is determined. The relative abundance versus mass-to-charge ratio that is produced from the substance by the mass spectrometer is called the mass spectrum of the substance.

Matrix-assisted laser desorption-ionization mass spectrometry (MALDI–MS) MALDI–MS technique enables ionization a broad range of materials, with high sensitivity, broad mass range of materials, fast sample preparation, and the absence of fragmentation.

Matrix isolation (MI) Technique used with Fourier transform infrared spectrometer (FT-IR) to detect spectroscopically transient chemical species at ambient temperatures.

Nuclear magnetic resonance spectroscopy Measures the absorption of light energy in the radio-frequency portion of the electromagnetic spectrum. ^{1}H NMR spectroscopy furnishes indirect information about the carbon skeleton of organic molecules. In ^{13}C NMR peaks corresponding to all carbon atoms are recorded.

Paper chromatography (PC) Technique in which the fixed phase is a sheet of filter paper. The sample is placed near the edge of the paper as a small spot. The edge is then dipped in the developing solvent (mobile phase). The solvent rises up the paper by capillary action taking the substance along. The positions of the spots are observed by visible or UV light or by spraying with a chromogen.

Raman spectroscopy (RS) Technique which involves the analysis of the intensity of Raman scattering of

monochromatic light as a function of frequency of the scattered light. The RS technique provides complementary information to infrared spectroscopy. Generally strong vibrations in the Raman absorption are weak in the infrared absorption.

Reaction GLC Certain chemical reactions of organic compounds (such as reduction, oxidation, dehydration) that take place in a gas–liquid chromatographic column. The products are detected by means of the conventional detectors.

Spectroscopy Instrumental method of assigning structural features and functional groups to organic or inorganic molecules; such features are displayed by intensities and patterns of spectroscopic signals.

Supercritical fluid chromatography (SFC) Technique in which the mobile phase (fluid) is maintained at temperatures somewhat above its critical point. The mobile phases used in SFC are gases such as freon, ethylene, or carbon dioxide. It has superior solution properties and enhances the chromatographic separation of higher molecular weight compounds. The column packings used in SFC are the same as those used in HPLC.

Supersonic jet spectroscopy (SJ) Method which measures with high discriminating power utilizing laser-induced or lamp-induced fluorescence spectroscopy.

Thermal chromatography Volatilization of organic compounds at high temperatures and their separation by chromatographic techniques such as TLC.

Thin-layer chromatography (TLC) Chromatographic technique in which the mobile phase is a liquid and the stationary phase is a thin-layer (usually 0.25 mm thick) of an adsorbent (silica gel, alumina, cellulose) spread homogeneously on a flat plate (usually a glass plate) of various dimensions.

Two-dimensional NMR spectroscopy (2D-NMR) Two-dimensional NMR refers to spectral data that are collected as a function of two time scales—evolution and detection.

Vacuumless spectrochemistry in the vacuum ultraviolet (VUV) Technique in which the emission wavelength region is below 200 nm. "Far ultraviolet" is more appropriate, since it includes access to this wavelength region by gas-filled optical techniques.

X-ray crystallography Method for determining the molecular structure of crystalline compounds which provides information on the positions of the individual atoms of a molecule, their interatomic distances, bond angles, and other features of molecular geometry.

COMPOUND DETECTION in organic chemistry refers to the methods of separation and identification of organic compounds. In modern technology this involves the use of chromatography (paper, thin-layer, gas–liquid, high-pressure liquid); spectroscopy (infrared, Raman ultraviolet and visible, nuclear magnetic resonance); mass spectrometry; and reaction chromatography (chemical reactions on thin-layer plates or gas chromatographic columns which can be carried out prior to, during, or immediately after the chromatographic separation). Pyrolysis and X-ray crystallography of organic compounds furnish important structural information on the partial structures or on the whole molecule, respectively. The combination (and computerization) of chemical, chromatographic, and spectroscopic techniques has become a more efficient tool for the detection and identification of organic compounds than any of these techniques.

I. INTRODUCTION

It was only about 40 years ago that chemists had the tedious task of identifying and characterizing unknown organic compounds especially in the area of natural products. This may involve degradation of the molecule followed by synthesis involving many steps. For example Woodward elucidated the structure of strychnine in 1947 and seven-years later successfully synthesized this compound.

The advent of computers and Fourier transform completely revolutionized the detection and identification of organic compounds. Modern automated instruments allow very small samples in the nanogram (10^{-9} g) range to be characterized in a very short time. The application of Fourier transform nuclear magnetic resonance (FTNMR) and Fourier transform infrared (FTIR) allows recovery of the sample in contrast to mass spectrometric (MS) determination which is a destructive but quite often necessary technique.

Modern methods especially in the separation of complex organic mixtures utilizing gas–liquid chromatography (GLC), high-pressure liquid chromatography (HPLC), and droplet counter-current (DCC) chromatography can separate samples rapidly and efficiently in the picogram range which until fairly recently has been impossible. Coupling the chromatographic instruments to spectrometers enables a partially automated analysis in even less time. The following coupling of chromatographic instruments has been performed: GC–MS, GC–FTIR, GC–MI–FTIR, GC–UV–VIS, HPLC–MS, HPLC–FTIR, HPLC–FTNMR and MS–MS. (Fig. 1).

These semi-automated systems of analyzing and characterizing small samples are vital to the natural product organic chemist and biochemist for detection of highly active substances in extremely low concentration in living organisms. A typical example is in the field of pheromones which includes insect sex attractants which differ quite

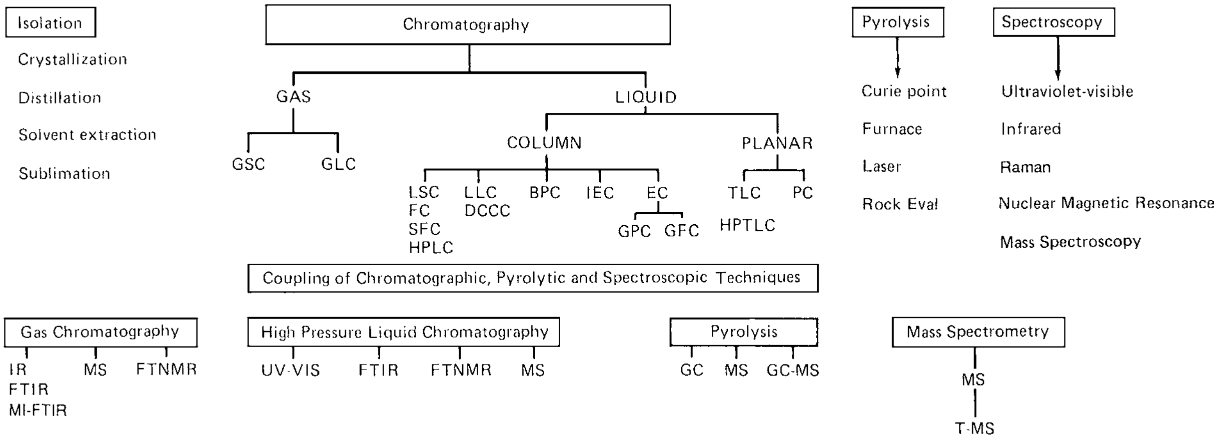


FIGURE 1 Chromatographic and spectroscopic techniques for detection and identification of organic compounds. GC, gas chromatography; GLC, gas–liquid chromatography; GSC, gas–solid chromatography; TLC, thin layer chromatography; HPTLC, high-performance thin layer chromatography; PC, paper chromatography; LSC, liquid–solid chromatography; FC, flash chromatography; SFC, supercritical fluid chromatography; LLC, liquid–liquid chromatography; DCCC, droplet counter current chromatography; PBC, bonded phase chromatography; HPLC, high pressure liquid chromatography; IEC, ion exchange chromatography; EC, exclusion chromatography; GPC, gel permeation chromatography; GFC, gel filtration chromatography; IR, infrared; UV, ultraviolet; NMR, nuclear magnetic resonance; MS, mass spectroscopy; FT, fourier transform; T-MS, Tandem mass spectroscopy; MI-FTIR, matrix isolation fourier transform infrared.

markedly in many insects. The concentration has often been found in the 10^{-9}–10^{-12} g range.

II. CHROMATOGRAPHIC METHODS

There is an old Dutch word for chemistry, *Scheikunde*, which literally means the art of separation. Indeed, separation methods from the chemical substance ultimately depends on separative operations.

Chromatographic methods occupy a rather unique position among modern methods in the field of detection, separation, and identification of organic compounds. The latest chromatographic methods provide simple techniques for separation, purification, and structure determination of organic compounds including the most complicated natural and synthetic macromolecules in biological and geological systems and their metabolic or breakdown products.

Chromatography permits the separation and partial description of substances whose presence is unknown or unsuspected. As an exploratory method, chromatography is indispensable in all sciences dealing with chemical substances and their reactions. Chromatography is, in fact, a physical method of separation in which mixtures are separated by distribution between two phases: a mobile phase and a stationary phase. The mobile phase can be a gas (as in gas chromatography) or a liquid (as in liquid chromatography).

Chromatographic methods can be classified according to the nature of the stationary phase. Gas chromatography is divided into GLC and gas–solid chromatography (GSC). Liquid chromatography is divided into two main types: column and planar chromatography methods, namely, thin-layer chromatography (TLC) and paper chromatography (PC). Column chromatography is subdivided further into five major column types which constitute HPLC. They are (1) liquid–solid chromatography (LSC), (2) liquid–liquid or partition chromatography (LLC), (3) bonded-phase chromatography (BPC), (4) ion exchange chromatography (IEC), and (5) exclusion chromatography (EC). The latter form includes gel permeation chromatography (GPC) and gel filtration chromatography (GFC).

GLC and HPLC are the most widely used techniques for separation of organic compounds. However, they are distinctly different techniques. For GC analysis the sample must be volatile and stable when the column is heated. Samples for liquid chromatography must be soluble in mobile phase and differentially retarded by the stationary phase. An active mobile phase increases the versatility of the liquid chromatographic technique. By contrast the mobile phase in GC is an inert gas which does not influence the separation. A variety of stationary phases (both polar and nonpolar) can be used. The versatility of liquid chromatography depends on the liquid phase and solid absorbent. Gases, liquids, and solids can be analyzed by GC. The normal range of molecular weights is from 2 to 500, although derivatives of carbohydrates ranging up to 1800 have been analyzed. Liquid chromatography is applicable to liquids and solids that are either ionic or covalently bonded. Molecular weights of 32 to 1,000,000 have been analyzed by LC. The flame ionization

detector in GC can readily detect 10^{-11} g and the electron capture detector can detect 10^{-10}–10^{-12} g of many organic compounds. The refractive index detector in liquid chromatography is limited to about 10^{-6} g; the ultraviolet (UV) detector can detect 10^{-8} g for many highly conjugated compounds. The resolution efficiency of a column is expressed in "theoretical plates." Thus, packed gas chromatographic columns of reasonable lengths can generate 2000–10,000 plates. Open tabular column (capillary size columns) can easily generate 50,000–100,000 plates. Total plates available in LC is greatly affected by pressure and analysis time available. Assuming 5,000 psi and flow rates of 0.5 ml/min, a reasonable upper limit is 10,000 plates.

Perfusion chromatography (P.C.) was recently developed by Fulton and his colleagues to exempt the user from choosing among speed, resolution, and binding in chromatographic separations of biomolecules. In P.C., transport into the particles occurs by a combination of convection and diffusion. The polymeric particles available under the trademark POROS contain two distinctive types of pores: (a) throughpores and (b) diffusive pores which are smaller and line the throughpores. Figure 2 shows a schematic diagram of conventional and HPLC diffusion chromatography packing particles, and Fig. 3 is a schematic diagram of P.C. packing particles, showing throughpores for rapid convective intraparticle mass transport and diffusive pores for high surface area and binding capacity. P.C. has been used in high-speed analysis and on-line monitoring using HPLC. The chromatogram of the separation of proteins (Fig. 4) shows that P.C. reduces run times to a few minutes (a tenfold decrease from conventional HPLC run times of 30–60 min) without loss of resolution. In biotechnological analyses of biomolecules,

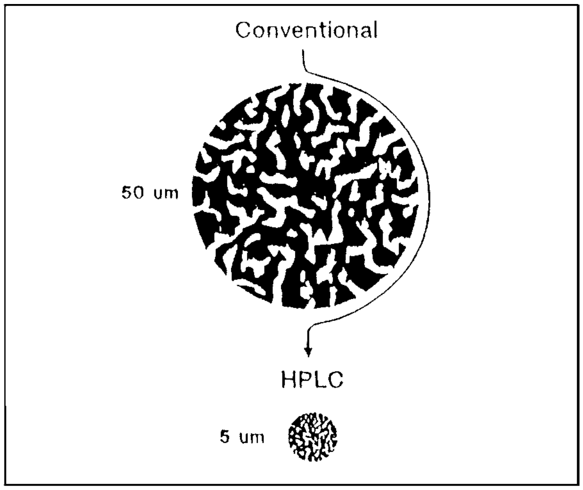


FIGURE 2 Schematic diagram of conventional and HPLC diffusion chromatography packing particles.

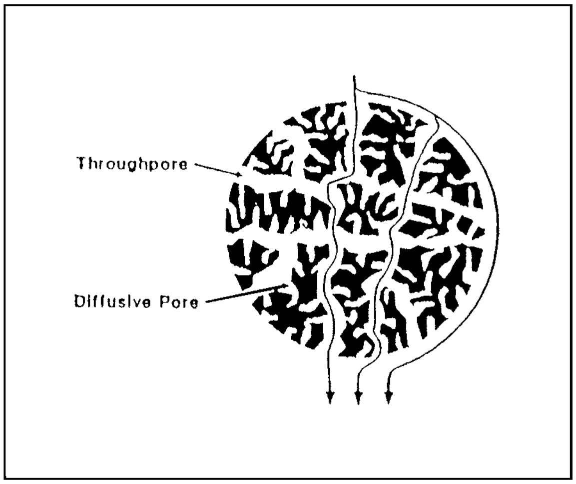


FIGURE 3 Schematic diagram of perfusion chromatography packing particles, showing throughpores for fast convective intraparticle mass transport and diffusive pores for high surface area and binding capacity.

P.C. can be used for on-line chromatographic monitoring of fermentation processes, primary recovery, and preparative chromatographic separations. P.C. has been used to scale-up to 600 times in the purification of the antibody IgG without loss of resolution. P.C. is a solution to the mass transport problem of liquid chromatography and allows

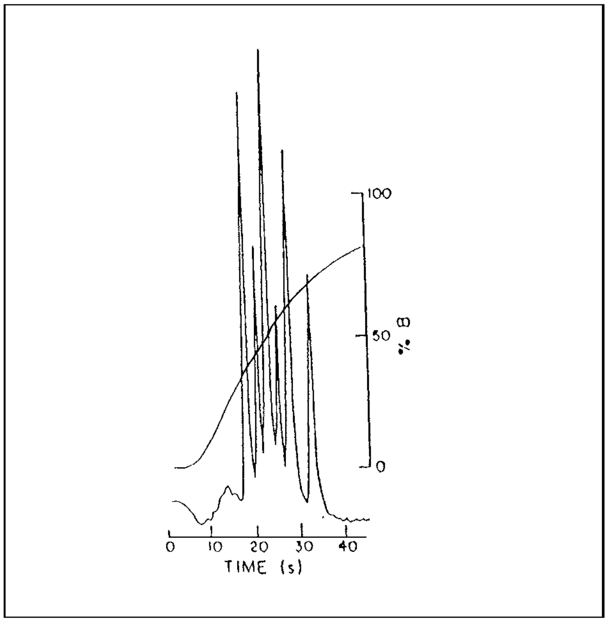
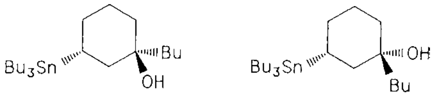


FIGURE 4 Analytical reversed-phase separation of standard test proteins on POROS perfusion packing. Column: 6 mmD/5 mmL POROS R/H. Sample: ribonuclease A, lysozyme, beta-lactoglobulins A and B, and ovalbumin. Mobile phase: 0.1% TFA in water. Gradient: 4–75% acetonitrile in 19 column volumes. Flow rate: 4.0 ml/min (850 cm/hr). Detection: OD 280 nm.

the use of 10- to 100-fold higher flow rates and lower cycle times without significant losses in either resolution or capacity compared with the conventional soft gels and HPLC columns.

III. FLASH CHROMATOGRAPHY

The traditional method for preparative separations of simple mixtures of organic compounds by means of column chromatography is giving way to flash chromatography (FC) or low-pressure liquid chromatography (LPLC) which was initially carried out by Still in 1978. FC is a simple, cheaper, and faster technique for routine purification of mixtures, allowing preparative separations from 10 mg to more than 25 g. The resolution associated with FC on a standard 5 in. high column of 40 μm silica gel is as great as 200 theoretical plates. This amount of separating power effectively means that compounds having R_f values on analytical TLC as close as 0.1 may be reliably separated with sample recovery of at least 90%. FC is basically an air-pressure-driven hybrid of medium-pressure and short-column chromatography which has been optimized for rapid separations. Elution of the components is extremely rapid, usually taking about 5 to 10 min. As with other forms of chromatography, sample size is one of the most important variables. The amount of sample that can be separated on a given column is proportional to its cross-sectional area and the degree of separation of the components as indicated by TLC. Still *et al.* successfully separated a 1-g mixture of epimeric alcohols **I** and **II** with only a 65 mg mixed fraction in 7 min on a 40-mm diameter column using 500 mL of 5% ethyl acetate/petroleum ether.



I II

Reversed-phase systems with nonpolar bonded C_8 and C_{18} silica are used in the separation of sugars. A mixture of 200 mg of fructose and sucrose were completely separated in the 3/1 acetonitrile/water mobile phase incorporating a silylamine bonded phase with silica in less than 1 h.

This method has been used for the separation of various products of organic synthesis and small biomolecules.

IV. DROPLET COUNTERCURRENT CHROMATOGRAPHY

The technique of countercurrent chromatography (CCC) has seen a rapid expansion following the introduction of new methods such as droplet countercurrent chromatography (DCCC), rotation locular countercurrent chromatography (RLCC), and coil planet centrifugation. These methods have the advantage of being more rapid and less solvent consuming than traditional CCC. Furthermore, the advent of commercially available, compact apparatus has led to a widespread acceptance of these new liquid–liquid techniques as standard laboratory procedures for the separation of natural products. The detection of compounds that are eluted from a DCCC can be performed by three methods: (1) UV detection for suitable UV-active substances, (2) monitoring of the fractions that are collected by TLC, and (3) weighing of fractions after evaporation of solvent. The majority of DCCC separations involve polar compounds, especially glycosides, which are often difficult to purify. Chloroform–methanol–water systems of varying compositions remain the most widely used, in view of the good formation of droplets and the convenient viscosity of this combination. The most notable developments in the application of DCCC have occurred in the field of polyphenols, in particular in the separation of tannins. DCCC has also been applied in the separation of natural products such as alkaloids, triterpene glycosides, steroid glycosides, basic steroid saponins, and glycosides of flavonoids. Rotation locular countercurrent chromatography (RLCC) relies on the percolation of one layer of a two-phase solvent system through compartments (loculi) that contain the second layer. During passage of the mobile phase, the loculi (connected into tubes) are constantly rotated, to increase contact between the two phases. Basically, RLCC has the same advantages as DCCC. As in DCCC, the apparatus can be run in either ascending or descending solvent modes but the formation of droplets is not a necessary condition of RLCC. Consequently, a broader range of solvent system is possible, and a system containing ethyl acetate (often incompatible with DCCC) has been used, for example, in the separation of flavonoids—an important application of this method has been the separation of enantiomers of ($\pm$)-norephedrine on an instrument consisting of 16 columns and each column containing 37 loculi. The stationary phase was sodium hexafluorophosphate solution at pH 4, and the mobile phase was (R, R)-di-nor-5-yl tartrate in 1,2-dichloroethane. Presumably, the enantiomers of (t)-norephedrine form different diastereotopic complexes with the tartrate ester, and these complexes are then partitioned differently between the two solvent phases. Separations by RLCC of a range of natural products, including flavones, xanthone glycosides, and antitumor antibiotics have been reported. RLCC provides a useful complementary method to DCCC in instances in which suitable solvent systems are not available.

Rotating Coil Method

The introduction of these potentially useful techniques is largely due to Ito in 1981. These methods involve CCD and eliminate the need for solid supports. One of the principal advantages is the speed of operation—the separation time (and consumption of solvent) approximating the level of HPLC in one instance involving the isolation of plant hormones by toroidal coil planet centrifugation.

No direct comparisons with DCCC or RLCC have been reported, and it will be of interest to see whether planet and toroidal coil centrifuges have the possibility of complementing or supplementing DCC or RLCC. Separation of amino acids and peptides have recently been reported utilizing the rotating coil method.

V. HIGH-PRESSURE LIQUID CHROMATOGRAPHY

Although the number of compounds detected by GLC could be increased by derivatization of polar functional groups, it has been suggested that only some 15% of all chemicals are capable of existing in the vapor phase.

HPLC has emerged as an instrumental technique offering rapid separations with simultaneous sensitive monitoring of the course of the analysis. HPLC in its most modern form is able to achieve separations in a matter of a few minutes which by previous techniques may have taken hours or days or may not have been possible.

Of particular importance in HPLC development has been the availability of specialized chromatographic column packings and sensitive on-line detection systems for continuous monitoring of the separations being carried out. These developments have led to systems which in favorable instances can on the one hand detect parts per billion (1 in 10^9) levels of organic compounds, and on the other hand be used for collecting gram quantities of pure chemicals by preparative HPLC. The lack of truly universal detectors has resulted in the development of several selective detectors as described in Table I.

Although LC detectors are sufficiently sensitive for trace analysis, they are nevertheless limited in their ability to provide adequate information for unequivocal identification. HPLC is particularly suitable in the separation of high-molecular-weight (up to 6 million) substances and thermally unstable biologically active products that cannot be volatilized without decomposition; aqueous and nonaqueous samples can be analyzed. HPLC has been applied to the analysis of natural and synthetic products such as amino acids, antibiotics, antioxidants, flavonoids, carotenoids, lipids, flavoring and aroma compounds, herbicides, pesticides, hormones, steroids, mycotoxins, polyaromatics, proteins, sugars, purines, dyes, vitamins, and water and air polutants. An example of HPLC applications includes the detection of the highly toxic aflatoxins sometimes found in peanuts, wheat, corn, and other grain crops. They are produced by fungus on the grain and are considered to be dangerous to human health when their concentration exceeds 20 ppb.

HPLC is an ideal system in pharmaceutical analysis, examination of drug formulation, degradation products of drugs that might be to toxic to humans, and detection of drugs in human fluids and tissues. HPLC is useful for polar and ionic compounds of medium to high molecular weight such as drugs and their metabolites.

A. Reversed-Phase HPLC (RP/HPLC)

This technique utilizes a nonpolar stationary phase, usually a fully porous microparticulate chemically bonded alkylsilica and a polar mobile phase are now recognized as the technique par excellence for the separation of polar, ionogenic solutes. For the separation of free amino acids, their derivatives, peptides, and proteins, the octyl (C_8) and octadecyl (C_{18}) hydrophobic phases bounded into 5- and 10-μm fully porous silicas with pore diameters in the range of 60–100 Å and 300–500 Å are being used.

B. Detection Systems

Photometric detectors with variable wavelength capability are most widely used currently in LC. As an example, the carcinogenic aflatoxins in cereal products absorb light strongly at both 254 and 365 nm. Many other compounds also absorb light in the former wavelength. At 365 nm,

TABLE I Comparison of Various LC Detectors

	UV	Refractive index	Fluorescent	Electrochemical	MS Fullscan	SIM[a]
Range of application	Selective	Universal	Very selective	Very selective	Universal	Very selective
Minimum detectable quantity, g.	10^{-9}	10^{-6}	10^{-12}	10^{-12}	10^{-9}	10^{-11}

however, most of the sample co-extractives are transparent and no longer interfere with the detection of aflatoxins. Although much work is perfomed with detectors that operate in the UV region of the spectrum a good deal is also practiced in the visible region.

C. Fluorescence Detection

Fluorimetry is well known for its very high selectivity and sensitivity to very small quantities of some samples occurring in biological fluids while being completely insensitive to many other materials such as durgs, vitamins, and steroids. GC detectors such as electrical conductivity and electron capture are also used in LC.

D. Radioactivity Detectors

There are many applications in the studies of the metabolism of drugs, pesticides, etc., in which radioactive samples are employed to enable the compounds of interest to be detected at very low concentrations. The important current trends in HPLC are the use of super critical fluids as eluants and the coupling of HPLC with NMR, MS, or FTIR.

An equally important trend is a reduction in column size and a concomitant increase in analytical speed. Short columns can reduce analysis time, and efficiency may be as high as 5000 theoretical plates since the particle size of the packing material is 3 μm. The analysis time is about 1 minute. A standard HPLC column has a diameter of 4.6 mm as compared to 2, 1, or even 0.5 mm for the new columns. The small size of these columns allows the use of more expensive solvents; deuterated solvents might be used when it is necessary to analyze the collected samples by NMR. The use of ultramicrobore (50-μm diameter) columns allows feeding of the eluted peaks directly into a mass spectrometer of FTIR instrument.

E. Preparative Liquid Chromatography

Most preparative HPLC columns are capable of separating as much as 10 g of sample at a time. Industry often utilizes columns 13 to 22 cm in diameter and 53 cm long having flow rates of 3 to 20 L/min. They can separate from 1 to 10 kg of sample per hour.

F. Chiral Separations Using Capillary GC

Chiral capillary chromatography has increasing utility in flavor, fragrance, pharmaceutical, and industrial chemical applications. The proper selection of a column is essential since chiral stationary phases illustrate different capabilities and applications. The compositions of commercial chiral phases vary by the size of cyclodextrin macromolecules and by substituents used to derivatize them. Both factors greatly affect a column's ability to resolve the enantiomers of different chiral compounds.

G. Enantiomeric Analysis of Amino Acids by HPLC

Most of the protein amino acids, except glycine, have at least one asymmetric (chiral) carbon atom and can exist as two isomers, designated D and L enantiomers. When the D and L enantiomers each react with a chiral (optically active) molecule of, for example, the L′ configuration, they form the diastereoisomeric compounds DL′ and LL′.

Amino acids of L configuration are more abundant in nature; the D-enantiomers are usually found in bacterial cell walls, antibiotic compounds, and rare biological molecules. The DL-form (the racemate) are found in geological specimens such as fossil shells and bones, or they are formed by abiotic synthesis.

Methods for resolving amino acids into their respective enantiomers are of importance in the preparation of peptides, drugs, and food additives as well as the extreme complex amino acid mixtures in meteorites in which the ready determination of optical purity is essential.

Two main approaches using HPLC are (1) covalent bonding of chiral ligands (which can complex copper (II) ions) to solid supports (such as polystyrene and polyacrylamide) and resolution of amino acids by eluting with a mobile phase containing copper (II) ions; (2) introduction of chirality into the mobile phase. Metal ions such as Cu(II), Zn(II), Co(II), and Mg(II), in conjunction with chiral ligands are added to the mobile phase. Thus, a Cu(II)–L proline complex as the chiral additive can be operated in conventional cation-exchange resin.

A good enantiomeric resolution of α-amino acids was recently achieved by using chiral complexes of copper (II) with *N,N*-di-*n*-propyl-L-alanine (DPA) as the additive in the mobile phase. Actually, the mixture of amino acids is separated into four groups by conventional ion-exchange chromatography and then resolved by means of the chiral DPA reagent (Fig. 5).

The relationship between temperature and enantiomeric separation may not always follow the same rules as conventional chromatography. Although the common liquid stationary phases used in GC do not possess adequate selectivity for enantiomeric separation, the addition of derivatized cyclodextrin macromolecules to common stationary phases often creates the ability to separate volatile enantiomers. Figure 6 shows that linalool and linalyl acetate have symmetrical peak shapes and excellent chiral separation on a Rt βDEXse at 25 mg per component on column.

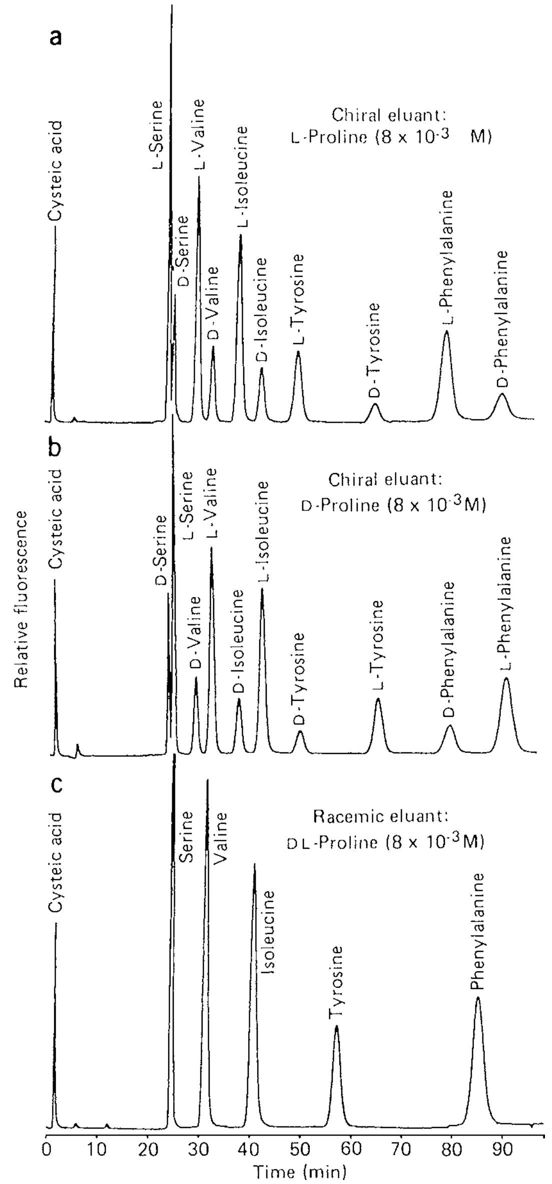


FIGURE 5 Effect of chiral eluant on the separation of D- and L-amino acid enantiomers by ligand-exchange chromatography.

H. Chiral Separations by HPLC

The recent developments in chiral separations by GC and LC has intensified the studies of asymmetrical (chiral) molecules in several scientific disciplines, such as biology, biochemistry, chemistry, and geochronology. Chiral separations are of great interest in pharmacology and pharmaceutical sciences because enantiomers can have different pharmacological and toxicological effects in humans. Different bioactivities have also been discovered for enantiomeric pesticides, insecticides, and herbicides. The chromatographic separation of enatiomers can be achieved in two ways. *Indirect separation* is based on the formation of diastereomeric derivatives with covalent bonds between the solute and an optically active reagent before chromatographic separation. The second approach, *direct separation*, is based on the formation of labile diastereomeric complexes with a chiral complexing agent (selector) in the stationary or mobile phase.

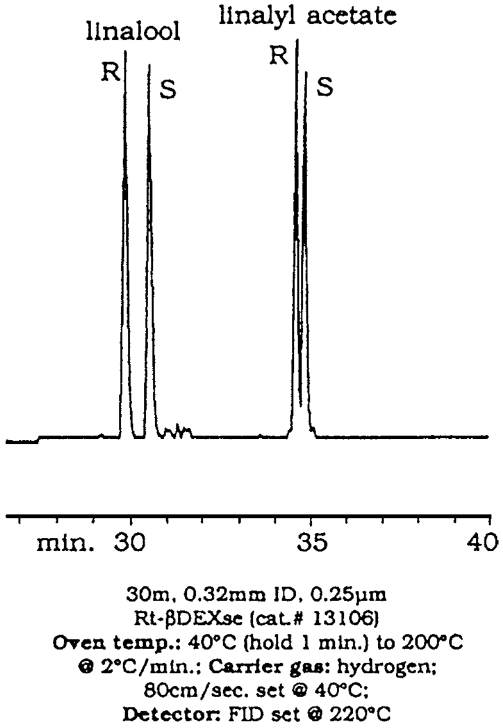


FIGURE 6

Some of the applications of the chiral stationary phases are for the determination of the enantiomeric purity in product control, confirmation of an asymmetric synthesis, and analysis of enantiomers found in biological fluids such as urine and blood. LC has been used to separate enantiomers in the milligram to gram scale. NMR spectroscopy using shift reagents has been used successfully in the determination of enantiomeric natural products such as monoterpenes.

VI. GAS CHROMATOGRAPHY

A. Derivatization Gas Chromatography

Conversion of sample compounds into volatile derivatives make it possible to separate and analyze by GC, groups of compounds for which GC analysis would otherwise be impossible, e.g., amino acids, sugars, prostaglandins, and related compounds. The presence of different polar

groups in the molecules of such parent compounds is the most significant source of difficulty associated with their GC analysis.

Carboxyl, hydroxyl, carbonyl, and amino groups, because of their polarity and tendency to form hydrogen bonds, are responsible for the low volatility of the compound and for other phenomena (such as strong absorption on the stationary phase) that make direct GC either difficult or impossible. The separation of closely related compounds is easier after their conversion into suitable derivatives. For example, the sensitive and labile prostaglandins such as $PGF_{2\alpha}$ are converted to their trimethylsilyl derivatives which are stable for GC analysis.

The combination of GC with MS (GC–MS) for which special derivatives are being developed, gives characteristic fragments that make identification and quantitative evaluation easier. Derivatives commonly used to protect these groups are usually less polar than the original groups. Efforts aimed at developing a single procedure in which several or all functional groups in the molecule could be converted into a suitable derivative in one reaction step led to the development of methods for the preparation of silyl, acyl, isopropyl, and other derivatives.

Esters are common derivatives of carboxyl groups. Methyl esters are the most often used as they have a sufficient volatility even for the chromatography of higher fatty acids contained in fats. A number of methods have been developed for their preparation, often exclusively for the purpose of GC determination.

Elegant methods for esterification with diazomethane and methanolic solutions of BF_3 or BCl_3 are fairly widespread. Reactions with methanol can also be catalyzed with HCl or H_2SO_4. Methyl esters can also be prepared by pyrolysis of tetramethylammonium salts in the inject port, and by esterification on an ion exchanger:

$$\text{R—COOH} + \text{CH}_2\text{=}\overset{+}{\text{N}}\text{=}\overset{-}{\text{N}} \rightarrow \text{R—COOCH}_3 + \text{N}_2\uparrow$$

$$\text{R—COOH} + (\text{CH}_3)_4\overset{+}{\text{N}}\text{OH}^- \rightarrow \text{RCOO}\overset{-\,+}{\text{N}}(\text{CH}_3)_4 \xrightarrow{\Delta} \text{RCOOCH}_3 + (\text{CH}_3)_3\text{N}\uparrow$$

Ethers are useful for protecting hydroxyl groups. Hydroxyl groups of high molecular weight compounds, such as sugars and sterols, are converted into ethers by reaction with methyl iodide in the presence of silver oxide.

$$2\text{ROH} + 2\text{CH}_3\text{I} + \text{Ag}_2\text{O} \rightarrow 2\text{R—OCH}_3 + 2\text{AgO} + \text{H}_2\text{O}$$

Aryl derivatives are common for hydroxy, amino and thiol groups.

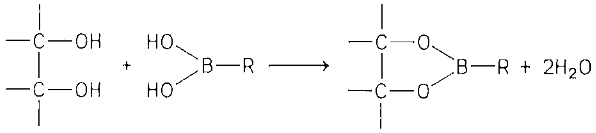
$$\begin{array}{lcl} \text{R—OH} & & \text{R—O—COR}' \\ \text{R—NH} & \xrightarrow[\text{R}'=\text{aryl}]{(\text{R}'\text{CO})_2\text{O}} & \text{R—NH—COR}' \\ \text{R—SH} & & \text{R—S—COR}' \end{array}$$

Silyl derivatives are probably the most commonly used for protecting functional groups of volatile substances in gas chromatography. Trimethylsilyl (TMS) derivatives can be prepared by the reaction of trimethylsilylating agents such as trimethylsilyl chloride with groups containing active hydrogen.

$$\begin{array}{rcl} \text{—OH} & & \text{—O—Si(CH}_3)_3 \\ \text{—COOH} & & \text{—COO—Si(CH}_3)_3 \\ \text{—SH} & \xrightarrow{\text{Me}_3\text{SiCl}} & \text{—S—Si(CH}_3)_3 \\ \text{—NH}_2 & & \text{—NH—Si(CH}_3)_3 \\ \text{=NH} & & \text{=N—Si(CH}_3)_3 \end{array}$$

1. Cyclic Derivatives

If two or more functional groups which should be protected occur in the molecule of the substrate, blocking can be accomplished with a bi-functional reagent, thus producing a cyclic product. Cyclic boronates, for example, can be used for the GC of compounds containing *cis*-diol groups in the 1,2- and 1,3-positions.



2. Separation of Enantiomers

Two basic approaches to the separation of enantiomers of amino acids have been applied: (1) amino acid derivatives are chromatographed on optically active stationary phases such as *N*-acyl alkyl esters, ureides or N-acyl alkyl esters of dipeptides and (2) GC separation is performed on conventional stationary phases and the derivatives of amino acids are prepared by reaction with optically active reagents. The bifunctionality of amino acids offers the use of either optically active esters or acyl derivatives.

3. Cyclodextrins

The cyclodextrins (cycloamyloses) are torus-shaped molecules consisting of 6, 7, 8, and 9 α-1,4-linked D-glucopyranose units, α, β, γ, and δ-cyclodextrin (α-CD [Fig. 7], β-CD, γ-CD, and δ-CD). Cyclodextrins have a hydrophobic center and a hydrophilic outer surface. In aqueous solution both the inner and the outer

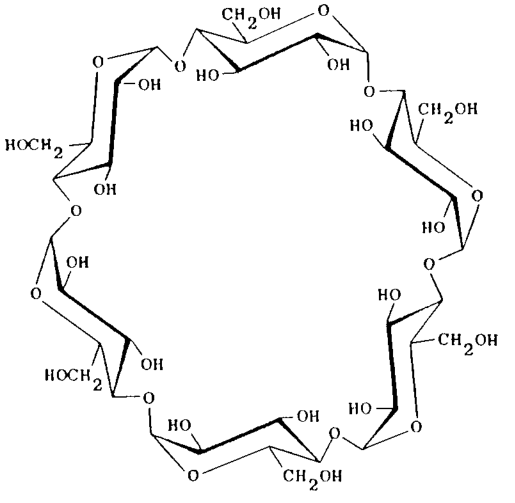


FIGURE 7 α-Cyclodextrin.

surfaces attract water, and the β-CD, for example, becomes distorted as a result of hydrogen bonding in the water molecules in the cyclodextrin cavity. Such hydrated molecules represent a high energy state which readily accepts a guest organic molecule in place of the water to form a complex. The reaction of cyclodextrins with bi- or polyfunctional compounds yields copolymers. The ability of the cyclodextrin cavity in these polymers to form complexes allows them to be used in chromatographic separations such as liquid, gel inclusion, affinity, and supercritical fluid chromatography.

a. Applications of cyclodextrins.

i. Chiral NMR shift reagents. α-CD and β-CD in D_2O, ^{19}F NMR spectra of the A_3B_3 type have been observed for $C_6H_5C(CF_3)_2OH$, the induced nonequivalence between CF_3 groups arising from guest accommodation in the optically active void of the host.

ii. Resolving of enantiomers. It was observed that complex formation is slightly stereoselective, and this property is used to resolve some enantiomers. The resolution of D,L-ethyl chloromandelate, sugars, penicillins, estrogens, and drugs such as suprofen were resolved using β-CD bonded phase chromatographic columns. Cyclodextrins have been used as stationary phases in capillary gas chromatography (GC) to separate optical isomers.

iii. Analysis of elements in biological materials. Recently, it has been reported that gallium (Ga) can be detected fluorometrically in biological materials, for example, liver, kidney, bone, and urine. Gallium forms an inclusion complex with benzyl-2-pyridylketone 2-pyridyl hydrazone-β-CD at a pH of 5 to form an intense yellow fluorescence. The detection and sensitivity limits of Ga were as low as 0.25 and 0.17 ng/mL, respectively.

4. Sugars and Related Compounds

The low volatility of sugars and related compounds such as amino sugars, polyalcohols, and aldonic and uronic acids, caused by the presence of several functional groups in one molecule as well as their high molecular weight and thermal lability are the reasons that they cannot be analyzed by GC methods. Another complication occurs in the formation of α- and β-anomers and pyranose and furanose rings, either during the preparation of the derivative or during the analysis itself.

The hydroxyl groups of saccharides and related sugars are silylated relatively easily; TMS derivatives have been widely applied to this group of substances.

5. Insecticides and Pesticides

This group of substances includes chemically very different compounds, such as ureas and carbamates, organic phosphorus and sulfur compounds, chlorinated hydrocarbons, and heterocyclic compounds. Depending on the type of functional groups in the molecule, common derivatives of these groups are applied. For the analysis of antibiotics TMS derivatives are generally prepared. Vitamins include compounds that differ considerably in their chemistry and therefore the range of possible derivatives is fairly wide.

6. Separation and Identification of Geometrical and Optical Isomers

Lipids are readily separated by chromatographic techniques into various components such as, for example, alcohols, sterols, fatty acids, and esters. However, these are not mostly single compounds but groups of compounds differing in chain length, degree of unsaturation, position of functional groups, and stereochemistry. The separation of some of these groups is accomplished by subjecting them to chromatography on adsorbents that have been impregnated with compounds having the ability to complex preferentially or interact with specific functional groups.

Unsaturated compounds such as fatty acids (oleic, linoleic, etc.) form π complexes: $AgNO_3$ TLC is used extensively to fractionate methyl esters of fatty acids according to their degree of unsaturation. Aromatic hydrocarbons as donors of π electrons are capable of forming donor–acceptor complexes with substances having electron-accepting properties, such as caffeine, tetracyanoethylene, and polynitro substances (e.g. 1,3,5-trinitrobenzene and 2,4,7-trinitrofluorenone).

Chromatographic resolution of optically active isomers requires the introduction of an asymmetric environment either intramolecularly by conversion to diastereomers or intermolecularly by the use of chiral stationary or mobile phases. In GC excellent resolution of derivatized amino

acids as been achieved with diastereomers, as well as with chiral stationary phases.

7. Identification Methods Based on Comparison of Retention Data

The most frequently used method for establishing the identity of an eluting component is comparison of the retention data of an "unknown" peak with the retention of a similar injection made under identical operating conditions of a reference substance, which, based on other considerations, e.g., a known synthesis precursor, could probably occur in the sample. Where possible the injection should be repeated with several chromatographic phase systems which exhibit different types of selectivity, i.e., a normal partition, a reversed-phase system, and a liquid–solid (adsorption) system. A vast amount of information exists relating to retention characteristics of samples in GC systems with chemical structure.

8. Identification Methods Using On-Line Selective Detectors

The most common application of this approach is the tentative identification of compounds that contain characteristic UV or visible absorption spectra. The use of more than one detector, linked in series or parallel after the chromatographic columns, can provide comparative information which reduces the possibility of incorrent assignment of the identity of a component. A simple example is the use of a UV absorbance detector in line with a differential refractive index (RI) detector. The latter will respond to most substances, whereas the former detector is quite selective in its response.

9. Monitoring of Column Effluents by MS

The combination of GC–MS and computerized data handling systems has proved to be one of the most powerful analytical methods for identifying minute (10^{-12} g) components which may be present in chemical samples. The greatest success has been in its application in the fields of forensic science, pollution, and biochemistry.

B. Reaction Gas Chromatography

Reaction GC is a variation of GC in which chemical reaction is coupled with the chromatographic separation. Chemical transformations in analytical reaction gas chromatography always take place in an integral chromatographic system, in a reaction syringe, a precolumn reactor, or the column itself. The combination of the chemical and the chromatographic methods is a more efficient tool for the identification of organic compounds than either of the two individual methods alone.

1. Esterification *in Situ*

Since organic acids are polar, they are converted into esters prior to analysis. The esterification is carried out either in front of the gas chromatographic column or directly on the column.

2. *In Situ* Hydrolysis

Hydrolysis is widely used in reaction GC for the identification of unstable and reactive compounds. A consecutive chromatographic separation, saponification, and chromatographic analysis of the resulting products was used for the identification of a mixture of high-boiling esters.

3. Dehydration and Decarboxylation

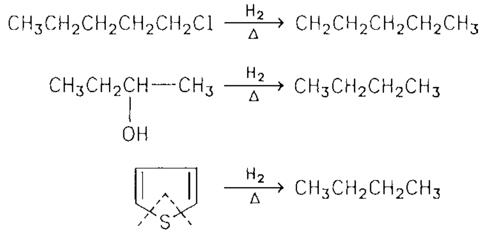
Monobasic organic acids are decarboxylated to hydrocarbons having one atom of carbon less than the acid. Alcohols are dehydrated and the olefins obtained are hydrogenated to the corresponding hydrocarbons. These transformations are carried out in a reactor at 250–300°C.

4. Carbon Skeleton Determination of Organic Compounds

The technique of carbon-skeleton chromatography is based on removing the functional groups from a compound and reducing the double and triple bonds. A hot tube containing a catalyst is introduced into the GC pathway and hydrogen is used as the carrier gas. As the injected compound passes over the hot catalyst, the compound is chemically degraded to its carbon skeleton. The hydrocarbon products pass into the GC and are identified by their retention times. In this technique hydrogenation, dehydrogenation (hydrogen abstraction), and hydrogenolysis (cleavage of functional groups or heteroatoms) may occur.

Hydrogenation, the saturation of multiple bonds, greately reduces the number of possibilities in determining the carbon skeleton. Dehydrogenation, the abstraction of hydrogen, takes place with cyclohexane derivatives and forms aromatic compounds at elevated temperatures of about 300–350°C. Hydrogenolysis involves the cleavage of functional groups from a molecule and the addition of a hydrogen atom to each of the cleaved ends.

In typical reactions (catalyst temperature 300°C, H_2 flow) the parent hydrocarbon is obtained from halides, alcohols, and heterocyclic compounds containing sulfur, oxygen, or nitrogen:

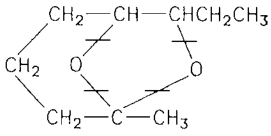


When an oxygen or nitrogen functional group is on the terminal carbon atom (aldehyde, primary alcohol, ester, ether, amine, amide, or carboxylic acid), the next lower homologue of the parent hydrocarbon is obtained, although the parent compound may be produced concurrently:

$$CH_3(CH_2)_{10}CHO \rightarrow CH_3(CH_2)_9CH_3$$

$$CH_3(CH_2)_{15}CH_2OH \rightarrow CH_3(CH_2)_{14}CH_3$$

Carbon-skeleton chromatography is not advanced as a quantitative procedure. Thus, carbon-skeleton chromatography structure was used to elucidate the structure of brevicomin,



brevicomin

an ingredient of the sex attractant of the western pine beetle. A large number of possible structures based on spectral evidence were narrowed down to a few based on the production of nonane in carbon-skeleton chromatography.

Hydrogenation is widely used to determine the structure of unsaturated compounds and to determine the olefin content of a wide variety of mixtures. The simplest means of hydrogenating microgram amounts of a sample "instantaneously" in the chromatographic pathway involves a hydrogenation catalyst and a hydrogen carrier gas. This technique was used for fatty acid methyl esters and a variety of unsaturated compounds (such as alcohols, amides, amines, ketones, esters, ethers, and nitriles) at the microgram level.

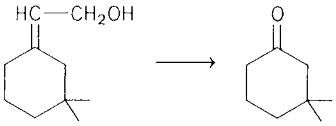
5. Locating Double Bond Position

Microozonolysis is used to determine the position of the double bonds. The aldehydes formed by passing ozone through a solution containing olefins is cleaved by triphenylphosphine to aldehydes or ketones (depending on whether the olefin is substituted):

$$RCH{=}CHR' \rightarrow RCHO + R'CHO$$

$$R'RC{=}CHR'' \rightarrow R\text{—}\overset{\overset{\displaystyle O}{\|}}{C}\text{—}R' + R''CHO$$

Thus, the structure of the following terpene alcohol, one of the components of the boll weevil sex attractant, was provided in part by microozonolysis when it yielded 3,3-dimethylcyclohexanone.



3,3-dimethylcyclohexanone

6. Selective Removal of Compounds from Mixtures

Compounds containing certain functional groups may be "subtracted" by specific chemical reagents in the gas chromatographic pathway. These effects are recognized by comparing chromatograms made with and without exposure to the chemical. Thus, primary and secondary alcohols are removed by boric acid, aldehydes and ketones by hydroxylamine, and carboxylic acids by zinc oxide. Subtractions are usually accomplished by including within the system a stainless steel loop consisting of the reactive chemical coated on an inert gas chromatographic support. The loops are most useful for determination of ozonolysis products. This inclusion of a loop containing 5% *o*-dianisidine on a gas-chromatographic support provides the information by subtracting aldehydes and allowing ketones to pass. Furthermore, $AgNO_3$ is used to hinder the passage of olefins.

A mixture of hydrocarbons can be separated by three selective adsorbents: molecular sieves which retain the *n*-paraffins, mercuric perchlorate which adsorbs the unsaturated compounds, and a stationary liquid phase which retains the aromatic hydrocarbons. Boric acid on a stationary phase such as Chromosorb P is used for the removal of alcohols from a mixture of organic compounds, the products are nonvolatile esters. A similar technique was used for removal of terpene alcohols from a mixture of terpenoids. Acids can be adsorbed on potassium hydroxide deposited on quartz powder. This technique was found to be suitable for the analysis of compounds having active hydrogens such as fluorene, indene, carbazole, indole, and pyrrole as well as steroids such as estrogens and ketosteroids. *o*-Dianisidine quantitatively subtracts aldehydes, ketones, and epoxides, and phosphoric acid subtracts epoxides.

The identification of the chemical type of compound can be carried out by dividing the gas chromatographic effluents into many different streams, each of which passes through a suitable color reagent as indicated in Table II.

The GC method is also being applied to the elementary analysis of carbon, hydrogen, oxygen, nitrogen, sulfur, and halogen organic compounds. This permits shorter

TABLE II Color Reagents Used for Detecting Gas Chromatographic Effluents

Class of compound	Reagent	Color
Alcohols	Potassium dichromate/nitric acid	Blue
Aldehydes	2,4-Dinitrophenylhydrazine	Yellow or red precipitate
Ketones	Schiff's reagent	Pink
	2,4-Dinitrophenylhydrazine	Yellow or red precipitate
Esters	Ferrous hydroxamate	Red
Mercaptans	Sodium nitroprusside	Red
	Isatin	Green
	Lead acetate	Yellow precipitate
Sulfides	Sodium nitroprusside	Red (primary)
	Sodium nitroprusside	Blue (secondary)
Nitriles	Ferrous hydroxamate/propylene glycol	Red
Aromatics	Formaldehyde/sulfuric acid	Red
Alkyl halides	Alcoholic silver nitrate	White precipitate

analysis time, reduction of sample weight, and also increases the accuracy of determination.

7. Combination of TLC and Vapor-Phase Chromatography

The combination of TLC and GLC enables a more complete separation and identification of complex mixtures of organic compounds. In most cases the mixture is subjected to preliminary separation by TLC followed by a complete separation of the preseparated compounds of the mixture by GC. The TLC–GLC technique has been applied in the analysis of citrus and other essential oils and for determination of steroids in urine as well as unsaturated hydroxy acid and glycerides which are separated on a thin layer of silica impregnated with $AgNO_3$ and then subjected to GC. The fractions from the chromatoplates can also be altered prior to running a vapor-phase chromatographic separation. Thus, acids are converted to methyl esters, alcohols and sterols to acetates, etc., prior to GLC. An application of the combination of TLC with GLC is the direct application of the compounds to the thin-layer plate as they emerge from the exit tube of the GC. This technique was used for the analysis of fatty acids, steroids, essential oils found in coffee essence, and many alkaloids. The association of TLC techniques with GC–MS was described for various separations and identifications of organic compounds.

C. Pyrolysis Gas Chromatography (Py/GC)

Prolysis is the thermal fission of naturally occurring and synthetic polymers producing a range of smaller molecules. Analysis of these products enables a profile of the original compound to be reconstructed. The pyrolytic technique is usually integrated with a gas chromatograph, a mass spectrometer, or both. This arrangement enables pyrolysis products to be analyzed immediately, so that transfer losses and secondary degradations are minimized. The resulting chromatogram or mass spectrum may provide qualitative information concerning the composition or identity of the sample, quantitative data on its constitution, or it may enable mechanistic and kinetic studies of thermal fragmentation processes.

The combined Py/GC/MS technique requires a minute amount of sample, and the analysis time is very short. Furthermore, such systems are now totally automated and computerized.

Analytical pyrolysis has been used successfully in many disciplines such as polymer chemistry, organic geochemistry, soil chemistry, forensic sciences, food science, environmental studies, microbiology, and extraterrestrial studies involving meteorites and lunar samples. A large number of organic substances found in nature are unsuitable for direct analysis by modern techniques such as column chromatography and mass spectrometry. This may be due to their complex structure and polar and nonvolatile character.

A significant step in the evolution of analytical pyrolysis was the combination of pyrolysis with a sophisticated physiochemical technique for the efficient separation and/or identification of the fragments. In 1959, a combined Py/GC system of polymers was introduced. Improvements in the analysis conditions were obtained by using the high-resolution capabilities of capillary columns. In 1970, a continued interest in the analysis of extraterrestrial samples such as lunar rocks and the Allende meterorite (found in Mexico) was reported, using the modified Py/GC/MS technique. Applying a pyrolysis/mass

spectrometry (Py/MS) technique such as the Curie point system offers considerable advantages: rapid analysis, automation of the system, and direct chemical information. Laser microprobe mass analysis (LAMMA) on milligram amount of sample has been reported. An alternative approach is the detection of very small (picogram) amounts of organic matter by direct ionization due to laser irradiation. Thus, digitonin (MW 1228) and sucrose (MW 342) were readily detected.

New techniques such as Py/GC/MS/DS (pyrolysis/gas chromatography/mass spectrometry–data system) and Py/MS/DS have been applied for detection and structural elucidation of complex organic compounds.

1. Pyrolysis of Synthetic Polymers

The thermal degradation of synthetic polymers has proven to be an extremely important analytical technique for revealing composition, structure, and stability profiles. It actually causes unzipping of the polymer chain to yield sequential monomer units. Thus, thermal depolymerization of rubber (at 700°C and atmospheric pressure) yielded isoprene and dipentene, whereas polystyrene yielded a series of monomeric and dimeric hydrocarbons.

2. Pyrolysis of Biological Molecules

The rapid extension of analytical pyrolysis into diverse fields such as taxonomy and soil chemistry has been largely due to progress in the pyrolysis of biological molecules. Such work has shown that the various classes of molecules with the possible exception of nucleic acids give highly characteristic pyrograms and has enabled the detection of the origin of fragments of complex samples such as bacteria and soil constituents. Interest in the pyrolysis of biological molecules is also due to the commercial importance of thermal degradation processes, such as flame retardation, use of biomass such as algae for production of oil and gas (energy), and thermal transformations of organic constituents of food during cooking. Thus, the sensitivity and specificity of Py/GC/MS has enabled octapeptides in infusion fluids and two closely related hormones, lypressin and felypressin to be detected in nanogram amounts in aqueous solutions.

3. Pyrolysis of Drugs and Natural Products

Py/GC, Py/GC/MS, and HPLC/MS are routinely used for the analysis of pure and formulated drugs or for the detection of active principles and metabolites in body fluids. The Py/GC method is used for detection of sulfonamides, barbiturates, and alkaloids such as morphine; heroine, phenacetin and caffeine (often used in forensic science).

The thermal reactions that accompany the smoking of tobacco and cannabis furnished important structural information on tobacco alkaloids. The pyrolysis of tobacco alkaloids yielded products such as quinoline and isoquinoline and nicotinonitrile which are probably derived from nornicotine and mysomine. Pyrolysis of natural polyenes such as β-carotene yielded ionene and small amounts of toluene, *m*-xylene, and 2,6-dimethylnaphthalene.

Although the identification and quantitation of drugs, e.g., narcotics and alcohol blood levels, are of forensic interest, the major impact of analytical pyrolysis in forensic science is in the identification of complex natural and synthetic substances such as blood, skin, hair, wood, soil, fibers, plastics, and waxes, which might be associated with a suspect and found in the scene of the crime.

4. Pyrolysis of Organic Geopolymers

Pyrolysis methods are used for detection and identification of natural organic polymeric products in the geosphere (such as humic substances and kerogens). The impetus for much of this work has been provided by the US space exploration program, when the efforts in taxonomy, biological molecules, and organic geochemistry which have resulted in Py/GC/MS studies, were undertaken on the surface of the planet Mars. The pyrolysis technique is also being used in the study of environmental pollutants (industrial processes, power stations, and fires). The use of the Py/GC/MS technique furnished important information on the structure of humic substances, melanoidins and coals.

The Rock–Eval method uses a special pyrolysis device (coupled with GC) in which a small sample of sedimentary rock (containing organic matter) is progressively heated to 550°C. During the assay the hydrocarbons already present in the rock (S_1) are first volatilized; pyrolysis of kerogen then results in the generation of hydrocarbons, hydrocarbon-like compounds (S_2), and oxygen containing volatiles, i.e., CO_2 (S_3) and water. Thus, $S_1 + S_2$, expressed in kilograms of hydrocarbons per ton of rock, is an evaluation of genetic potential (abundance and type of organic matter) of the rock.

5. Extraterrestrial Matter

Coupled pyrolysis techniques have been used for the analysis of extraterrestrial organic matter to provide evidence on molecular evolution within the solar system. Samples analyzed so far included meteorites collected on the surface of the earth, lunar rocks transported to earth and certain soils analyzed on the surface of that planet. The fully automated Py/GC/MS/DS analysis of Martian soils revealed only the presence of water and CO.

VII. SUPERCRITICAL FLUID CHROMATOGRAPHY

In supercritical fluid chromatography (SFC), the mobile phase is maintained at a temperature somewhat above its critical point. Since the physical properties of a substance in the supercritical state near the critical point are intermediate between those of liquids and gases at ambient conditions, it is designated as a fluid. For chromatographic purposes such a fluid has more desirable transport properties than a liquid. SFC is superior to LC in separating efficiency and speed. In comparison to a gas, a fluid shows about a 1000-fold increase in solution capabilities. The resultant enhancement in the migration rate of solutes is especially valuable in the analysis of higher-molecular-weight compounds. Furthermore, some ionic solutes are soluble in a supercritical fluid. This suggests that SFC may be applicable to the analysis of compounds such as the phospholipids, which cannot be volatalized for GC without decomposition.

The number of compounds that can be analyzed by SFC is potentially enormous. Out of the 10^6 known compounds which are currently more or less well characterized, only about 15% can be volatalized without decomposition. Compounds such as proteins, synthetic and natural polymers, lipids, carbohydrates, vitamins, synthetic drugs, and metal organic compounds may well be analyzed by SFC.

When a liquid and its vapor in equilibrium with each other are heated in a confined space, the intensive properties of the two coexisting phases become increasingly similar until, at the critical temperature, the two phases coalesce into a *fluid* and acquire the same properties. When this substance is heated beyond the critical temperature, a supercritical phase is obtained, the substance is then called supercritical fluid.

For comparative purposes some physical properties of a gas, a liquid, and a supercritical fluid are shown in Table III. The data in Table III show that the viscosity of a supercritical fluid is comparable to that of a gas and its diffusibility is between that of a gas and a liquid.

Any compound which is thermally stable to somewhat beyond its critical point can theoretically be used as the mobile phase. The mobile phases used in SFC include freons, ethylene, pentane, hexane, isopropanol, and carbon dioxide. It has been observed that the solubility of various solutes increases in supercritical phases. The column packings that are used in SFC are essentially the same as those used in HPLC.

The instrumentation in SFC is quite similar to that of modern high-resolution liquid chromatography. A number of modifications are made for LC to be suitable for operation with a supercritical fluid. The separation column can be coupled to online detectors other than UV, including MS, FTIR, FID, and other GC detectors.

Carbon dioxide offers many advantages; it is inexpensive, available in high purity, and innocuous. Its near-ambient critical temperature makes it attractive for use with thermally labile compounds. The UV absorbance of CO_2 is minimal, thus allowing spectrometric detection down to 190 nm.

The advantages of SFC are high resolution per unit time, orthogonal column selectivity compared to GC and HPLC, ease of fraction collection, and analysis of thermally labile molecules. It is likely that 20–40% of the solutes presently separated by HPLC are amenable to SFC separation with supercritical CO_2. It is possible that most separations carried out by normal-phase HPLC may be handled with good advantage by SFC.

Recent developments in SFC have not been inspired by any revolutionary new ideas. There has been progress in two directions in the area of SFC. One direction parallels the development of capillary GC, the second direction is similar to the development in HPLC. Because capillary SFC is a high efficiency separation method, it is most useful for the separation of complex mixtures. In combination with universal (FID) detection, it can be used as a "fingerprint" technique.

Packed column SFC with particles of 3 to 10 μm can be used with mobile phases of relatively high densities. At least 20,000 theoretical plates can be achieved within a few minutes on a routine basis.

A further advantage of SFC with respect to LC and GC is the separation of nonvolatile components and to introduce them to a mass spectrometer in such a way that both molecular weight information and reproducibly fragmented spectra can be obtained.

TABLE III

Property	Units	Gas	Liquid	Supercritical fluid
Density	g/ml	10^{-3}	1	0.3
Diffusibility	cm^3/sec	10^{-1}	$5 \cdot 10^{-6}$	10^{-3}
Dynamic viscosity	poise (g/cm sec)	10^{-4}	10^{-2}	10^{-4}

A. Interfacing Supercritical Fluid Chromatography with Fourier Transform IR (SFC–FTIR)

Supercritical fluids have different coefficients which are about 100 times greater than those of liquids and with very low viscosities. When open tubular columns are used in SFC, overall efficiencies on the order of 10^5 to 10^6 theoretical plates are obtained in reasonable analysis times.

FTIR is a powerful and highly specific detection technique. The combination of SFC with FTIR has tremendous scope because of the possibility of separating and identifying compounds which are not amenable to GC–FTIR analysis.

SFC–FTIR is a very useful technique in characterizing chemical additives in polymers. Many of these additives, which include slip agents, plasticizers, UV absorbers, light stabilizers and antioxidants are thermally labile or have a high molecular mass which makes them difficult to analyze by other methods.

The biological or pharmacological activity and effectiveness of chiral molecules depend largely on their configuration. Often only one of the enantiomers is pharmacologically beneficially active, while the other enantiomer may not only be inactive but also toxic. Most of the chiral stationary phases for TLC, GC, and LC of the chiral-cavity type act by allowing selective occlusion or intercalation of one enantiomer into chiral cavities in the phase matrix.

B. Superficial Fluid Chromatography (SCF) and SCF/MS for Analysis of Complex Hydrocarbon Mixtures

The physical properties of supercritical fluid mobile phases allow chromatographic separations of many compounds that are generally separated by LC. The density of a supercritical fluid is typically 100 to 1000 times greater than that of a gas, which imparts "liquid-like" solvating properties. However, the viscosities and diffusivities of supercritical fluids remain similar to those of gases. Consequently, high-solvating power mobile phases that behave similarly to gases define the advantages of SFC. These properties also allow successful interfacing to MS. The immediate application of SFC/MS in complex hydrocarbon analysis is for mixtures containing nonpolar to moderately polar compounds that are thermally labile or of relatively high molecular weight.

VIII. THIN-LAYER CHROMATOGRAPHY

Thin-layer chromatography (TLC) is a subdivision of liquid chromatography (LC) in which the mobile phase is a liquid and the stationary phase is situated as a thin layer on the surface of a flat plate. TLC is sometimes grouped with paper chromatography under the term *planar liquid chromatography* because of the flat geometry of the paper or layer stationary phases. TLC is a simple, rapid, versatile, sensitive, inexpensive analytical technique for the separation of organic substances. Since numerous stationary phases (sorbents) are available, such as silica, cellulose, alumina, polyamides and ion-exchangers, considerable versatility is available in the type of substances that can be separated.

TLC is a microanalytical procedure and provides for separations and at least tentative identification of substances in the milligram microgram, nonogram, and even picogram (pg) range. Adsorption TLC is very sensitive to differences in configuration that affect the free energy of adsorption onto the layer surface and is, therefore, well suited to the separation of structural isomers. Quantitative estimation of the separated compounds is carried out *in situ* by densitometric estimation of the TLC plates.

TLC has recently been transformed by the introduction of new and better plates, accurate and precise autospotters, new development techniques, and rapid scanning densitometers into a sophisticated, instrumentalized, and quantitative technique.

With respect to the TLC plates, there are two major advances. The first was the introduction to high performance TLC (HPLC) plates produced from silica gel of between 5 and 10 μm compared to the 12 to 25 μm commonly used for standard TLC plates. This resulted in more rapid and efficient chromatographic separation and reduced bond broadening, thus producing higher sensitivity. Second were a large number of bond-phases, of the type originally produced for HPLC, such as C_2, C_8, C_{12}, C_{18}, aminopropyl, cyanopropyl, diphenyl, and the so called chiral phase which are used in TLC.

The chiral plates are used extensively for the enantiomeric separation of racemic amino acids and their derivatives by means of a ligand exchange mechanism.

Detection and quantitation of analysis on the TLC plates have progressed considerably in recent years. A range of microprocessor-controlled scanning densitometers is now available, capable of providing UV and fluoresence detection at the nanogram level. Chromatographic processes have been reported to be filmed on video.

A. High-Performance Thin-Layer Chromatography

High-performance TLC enables us to carry out the most complicated separations. The HPTLC plates are prepared from optimized (e.g., particle size and particle size distribution) adsorbent layers and extremely even surfaces. The HPTLC plates offer greater separating efficiency (plate number and resolution) through smaller plate heights than the conventional TLC plates. Shorter analysis time, detection limits in the nanogram range with UV adsorption detection and in the picogram range with fluorometric detection are additional advantages.

The HPTLC plates may typically deliver some 4000 theoretical plates over a distance of 3 cm in 10 min. This

compares to typical values for conventional TLC plates of some 2000 theoretical plates over 12 cm in 25 min.

B. Impregnated Layers

The range of applications of both TLC and HPTLC is considerably expanded by means of different impregnation agents such as acids, bases, or salts added to layers in various concentrations. Stable hydrophilic stationary phases are formed by treatment with agents such as formamide, DMF, ethylene glycol, and various buffers. Lipophilic stationary phases for reversed-phase TLC are obtained by impregnation with liquid paraffin, undecane, and mineral and silicone oils. Impregnation with specific reagents aids the separation of certain types of compounds such as $AgNO_3$ for compounds with double bonds, boric acid or sodium arsenite for vicinal dihydroxy isomers; sodium bisulfite for carbonyl compounds, and trinitrobenzene or picric acid for polynuclear aromatic compounds.

C. Detection and Visualization

Following development, chromatograms are removed from the chamber and are air- or oven-dried to remove the mobile phase, zones are detected by various means. Colored substances may be viewed in daylight without any treatment. Detection of colorless substances is simplest if compounds show self-absorption in the short-wave ultraviolet (UV) region (254 nm) or if they can be excited to produce fluorescence by short-wave and/or by long-wave (365 nm) UV radiation. Otherwise, detection can be achieved by means of chromogenic reagents (producing colored zones), fluorogenic reagents (producing fluorescent zones), or by biological enzymatic methods.

Enzymatic reactions can be monitored on the plate, and the end products can be detected. Biological test procedures are used in the specific detection of biologically active compounds. Thus, detection of hemolyzing compounds such as saponins is achieved by casting a blood–gelatin suspension on the layer and observing hemolytic zones that are transparent and nearly colorless on the turbid red gelatin layer background. Another means of detection is the use of Geiger or flow counters or other specialized means to locate radioactive solutes.

Detection reagents may be impregnated into the layer prior to sample application and development. Chromogenic reagents are of two types: (1) general reagents that react with a wide variety of different compound types and can totally characterize an unknown sample, and (2) specific reagents that indicate the presence of a particular compound or functional group. The universal detection reagent iodine can be used as a 1% alcoholic solution spray, but more frequently, the plate is simply placed in a closed container containing a few iodine crystals. The iodine vapor forms weak charge-transfer complexes with most organic compounds which show up as brown spots on a pale yellow background within a few minutes. Sensitivities in the 0.1–0.5-μg range are often obtained with iodine.

Charring reagents (H_2SO_4) are suitable for glass-backed layers with inorganic (e.g., gypsum) binders only. Many charring reagents produce colored zones when heating is carried out at relatively low temperature; they form black zones at higher temperatures.

Spraying of a chromatogram with a 5% solution of phosphomolybdic acid followed by a brief heating at 110°C gives dark blue spots against a yellow background with a large variety of organic compounds.

A solution of Rhodamine B produces violet spots on a pink background. Antimony trichloride or pentachloride solution in carbon tetrachloride produce spots of different characteristic colors with many organic compounds.

Over 300 spray reagents are known to react more or less specifically with different functional groups to reveal natural products and organic or biochemicals as colored or fluorescent zones. Table IV contains a selection of specific detection reagents. Methods for the quantitation of thin-layer chromatograms can be divided into two categories. In the first, solutes are assayed directly on the layer, either by visual comparison, area measurement or densitometry. In the second, solutes are eluted from the sorbent before being examined further.

D. Radiochemical Techniques

Radioactive isotopes are widely used as tracers or labels for substances separated by TLC for following the causes of chemical and biochemical reactions, determining the distribution of substances in a reaction mixture, elucidating metabolic pathways of drugs, pesticides, pollutants, and natural substances in human, animal and plant tissues, and assessing the purity of isotopes. The most widely used labelled substances in TLRC contain ^{3}H (tritium) and ^{14}C.

The detection of radioactive substances on TLC plates is carried out by liquid scintillation, film registration or autoradiography, and by direct scanning. A thermomicro procedure for rapid extraction and direct application in TLC is the thermomicro application of substances (TAS) method.

The TAS method is a procedure for the isolation and separation of many substances from solid materials and their direct transfer to the starting line on a TLC plate. This is performed by application of heat, either by distillation

TABLE IV Detection Reagents for Different Functional Groups

Compound class	Reagent	Color
Alcohols	Ceric ammonium sulfate	Brown spots on yellow background
Aldehydes	2,4-Dinitrophenylhydrazine	Yellow to red spots on pale orange background
Ketones	Dragendorff–Munier modification reagent	Orange spots
Alkaloids	Iodoplatinate	Basic drugs yield blue or blue-violet spots
Amides	Hydroxylamine–ferric nitrate	Various colors on white background
Amines and amino acids	Ninhydrin	Yellow-pink-red or violet spots on white background
Carbohydrates	*p*-Anisaldehyde	Blue-green and violet spots
Carboxylic acids	Cresol green	Yellow color on blue background
Bases	Cresol green	Blue spots on green background
Chlorinated hydrocarbons and chlorine containing particles	Silver nitrate	Gray spots on colorless background
Ethanolamines	Benzoquinone	Red spots on pale background
Heterocyclic oxygen compounds	Aluminum chloride	Flavonoids produce yellow fluorescent spots
Hydrocarbons	Tetracyanoethylene	Aromatic hydrocarbons yield various colors
Hydroxamic acids	Ferric chloride	Red spots on colored background
Indoles	Ehrlich reagent	Indoles: purple. Hydroxy-indoles blue
Feroxides	Ferrous thiocyanate	Red-brown spots on pale background
Phenols	4-Amino antipyrine	Red, orange, or pink spots on pale background
Polynuclear aromatic hydrocarbons	Formaldehyde–sulfuric acid	Various colors on white background
Steroids	*p*-Toluenesulfonic acid	Fluorescence
Steroid glycosides	Trichloroacetic acid–chloramine T	Digitalis glycosides: blue spots
Terpenes	Diphenyl-phenyl-hydrazyl	Yellow spots on purple background
Vitamins	Iodine-starch	Ascorbic acid; white on blue background

or sublimation. The emerging vapors are deposited as a spot on the TLC-plate which is then chromatographed in the usual way. The spot can also be scraped off, extracted and analyzed by GLC.

This technique is useful in the fields of drugs, phytochemistry, food additives, and other natural and synthetic organic compounds. Many organic constituents of drug-containing plants were detected by this technique. Examples are constituents of essential oils, purines (e.g., caffeine), narcotics from drugs or tobacco, marijuana constituents, mescaline from Mexican narcotic fungi, organic contituents of fossil fuels, and oil-shales.

E. Coupled TLC Techniques

1. TLC–GC

There are a number of ways in which TLC can be combined to advantage with GLC. The spots obtained from TLC may be eluted, concentrated, and then subjected to GLC analysis. This method has been used in the analyses of lipids, steroids, alcohols, fatty acids, esters, glycerides, hydrocarbons, essential oils, and many other natural and synthetic organic compounds. Methyl esters of fatty acids are first separated on silver-nitrate-impregnated silica gel layers according to the degree of unsaturation. The separated fractions are eluted and the products subjected to GLC separation. Furthermore, the resolved methyl esters may be treated to reductive ozonolysis and the resulting fragments analyzed by GC, thus furnishing important structural information.

2. GC–TLC

A recent application of the combination of TLC with GLC has been the direct application of the compounds to the thin-layer plate as they emerge from the exit tube of the gas chromatographic apparatus.

It should be pointed out that GLC separates according to the relative volatility, and TLC separates according to the functional groups present. A coupling device was constructed in which the thin layer plate is moved logarithmically with time, while the gas chromatogram operates under isothermal conditions.

3. *In Situ* Chemical Reactions on TL Plates

In order to detect and identify minute quantities of organic compounds the sample is spotted on a plate which is then

covered with a reagent. After a very short time, (mostly a few seconds or minutes) the plate is developed in a suitable solvent system whereupon the reaction products are separated and identified by color or by spectroscopic methods. A great deal of information can be gained from the *in situ* reaction technique with the expenditure of very small amounts of material. The following *in situ* reactions on TL plates have been carried out so far: acetylation, dehydration, formation of derivatives: (acetates, dinitrobenzoates, DNPS, methyl esters, phenylisocyanates, and semicarbazones), diazotization, esterification, halogenation, catalytic hydrogenation, acid and alkaline hydrolysis, isomerization, nitration, oxidation, photochemical reactions, reduction, and Diels–Alder reactions.

F. Lab on-a-Chip

A Lab on-a-chip is a planar device on which or in which a number of chemical processes are being performed in order to go from reactants to products or from a sample to analysis. There is a distinction between microfluidic devices and microarray devices and DNA analysis.

DNA microchip arrays contain from several hundred to several thousand immobilized DNA reagents. They provide a systematic way to survey DNA and RNA variation and may well provide a standard tool for molecular biology research and clinical diagnostics. The peripherals such as MS and fluorescence detectors sit off-chip and are typically benchtop instrument sized.

At present, the typical commercially produced microchip used for miniaturized chemical systems consists of a 2- or 3-cm-square sliver of silicon, glass, quartz, or plastic that is 50 μm. The chip is covered with a plate to contain the samples and reagents. At present it is possible to inject volumes as low as 1 pL onto chips with miniaturized chemical systems.

For its operation, a microfluidic chemical device requires a number of microcomponents, such as microfilter, microreactors, and microseparation columns, depending on the application. It also needs some means of driving the fluids around the chip. Recently, a microtechnology for producing "bioprocessors-on-a-chip" was developed. Such chips can be described as microfactories that can within minutes diagnose infections or diseases by separating bacteria, fatal cells, or cancer cells from blood. This technology relies on dielectrophoresis for separating bioparticles, such as cells, viruses, bacteria, proteins, and DNA with a high degree of selectivity and sensitivity.

With improvement in layout and packaging of microchips, it might even be feasible to integrate the detection and computer circuiting on the chip.

It should be pointed out that typical measurements in a laboratory take several hours to perform, but many of the desired applications of on-chip analytical systems require data in seconds, or at most a few minutes. Fast analysis is therefore an important aspect of miniaturization.

IX. FIELD-FLOW FRACTIONATION (FFF)

FFF is a chromatographic-like separation technique which is designed for fractionation of macromolecules, colloids, and particles, primarily on an analytical scale. While the chromatographic separation is achieved by differential distribution of species between one stationary phase and one moving phase. FFF uses only one moving phase. However, there are velocity difference within the moving phase, and separation is achieved by differential distribution of species between the regions of different velocity. FFF has broad application.

Sedimentation FFF has been used for the fractionation of polystyrene, latex beads, emulsions, artificial blood, viruses, and aqueous colloids, liposomes, albumin spheres, and DNA. Thermal FFF has been applied to different types of synthetic polymers. A recent development in FFF has been the increase of separation speed to allow fractionation on a minute scale instead of hours.

X. ION CHROMATOGRAPHY (IC)

lon chromatography (IC) usually contains an ion exchanger or 'pseudo ion exchanger'. The ion exchangers are derived from cross-linked organic polymers by affixing to the polymer ionogenic groups that are the source or the vital ion exchange processes. The IC is used for the determination for both inorganic or organic compounds and even species as 'non-ionic' as carbohydrates.

The determination of ions in environments as diverse as brine, milk, diesel exhaust, plating baths, urine and foodstuffs is a measure of its versatility. Improvements in resins and other stationary phases and in the sensitivity and automation of IC techniques should help to sustain the growth in this new area of analytical technology.

XI. CAPILLARY ELECTROCHROMATOGRAPHY (CEC)

CEC may be regarded as a combination of high-performance capillary electrophoresis (HPCE) and HPLC—a hybrid technique combining the best of both worlds.

CEC is a derivative of HPCE, wherein a packed bed capillary is used. The essential difference between HPCE

and CEC is in the nature of the separation. If the separation is achieved on the basis of differential partitioning between two phases, then it is chromatography. If it uses both, it is referred to as electrochromatography.

Compared with HPLC, CEC uses electroosmotic flow (EOF) rather than high pressure to force the mobile phase through the capillary. The advantage of EOF for CEC is that the flow profile is flat and thus precludes band broadening by trans-channel of radial diffusion. The result is that the number of plates is at least double that of HPLC. Another advantage of CEC is that the packing particles are smaller than those of today's HPLC systems.

Since there is no back pressure resisting the pump's ability to deliver the solvent at a given flow rate, as in HPLC, capillaries can be made longer. Thus, it is practical to utilize packings as small as 1 μm. These can be porous or nonporous, spherical or irregular, coated or uncoated, pure silica or mixed mode (silica/polymer, silica/alumina, etc.). The combination of longer capillaries packed with smaller-diameter particles makes it possible to obtain and use columns with 100,000–500,000 and perhaps even 1,000,000 plates per meter.

In conventional HPLC, achieving more than about 25,000 plates per meter requires coupling columns and long analysis times.

Using small internal diameter columns and small particle diameters it is possible to realize somewhat higher plate counts. It took 30 years, but today, capillary columns account for nearly 75% of the GC separations.

XII. SPECTROSCOPIC METHODS

There has been a profound impact on the application of spectroscopic techniques in the detection of organic compounds especially in the areas of natural products and polymers. The commercially available Fourier transform signal handling techniques are rapidly revolutionizing nuclear magnetic resonance (FTNMR), infrared spectroscopy (FTIR), and Raman spectroscopy (FTRS), as well as prototypes of mass spectrometry (FTMS) which are not yet available commercially. Previously organic chemists usually required all the spectroscopic data (UV, IR, NMR, and MS) to supplement microanalysis and general physical characteristics such as refractive index, density, and melting and boiling points in order to elucidate the structure of the organic compound. Furthermore, it was found that at least 50 mg of the material was needed in order to determine the structure. It is now possible with the modern available techniques to determine structures with less than a milligram of material.

A. Infrared Spectroscopy

Infrared (IR) spectroscopy is probably the quickest and cheapest of the spectroscopic techniques in determining the functional groups of the sample. The samples can be solids, liquids, or gases and can be measured in solution or as neat liquids mulled with KBr or mineral oil. Comparison of IR spectra of substances of known structure has led to many correlations between wavelength (or frequency) of IR absorption and features of molecular structure. Certain structural features can easily be established. For example, in an organic compound that contains only C, H and O, the oxygen can only be present as C═O, O—H, or C—O—C or a combination of these, such as the ester or carboxylic acid group.

The presence or absence of absorption in the carbonyl region (1730–1670 cm^{-1}) or hydroxy region (3700–3300 cm^{-1}) can serve to eliminate or establish some of these possibilities. One simple application of IR is to determine whether two samples are identical. If the samples are the same, their IR spectra (obtained under identical conditions) must be the same. If the two samples are both pure substances very similar in structure, the differences in the spectra may be so small that it would not be easy to detect them; it may even be beyond the power of the instrument to detect them. The absorption peaks of the spectrum of an impure sample is usually less intense than those of a pure sample and the spectrum will show additional peaks. The IR spectra of enantiomers are identical but those of diastereoisomers are different.

Simple mixtures can be determined from their IR spectra. For example mixtures of cycloalkanones show characteristic differences of the carbonyl group: cyclobutanone (1788 cm^{-1}), cyclopentanone (1746 cm^{-1}), and cyclohexanone (1718 cm^{-1}). IR spectroscopy has been applied to deuteration of organic samples. It is found that the C—H stretch bands disappear from the 3000 cm^{-1} and the C—D stretch bands appear in the 2200 cm^{-1} region. It is possible by calibrating pure compounds to determine the isotopic purity in routine analysis.

Modern IR analysis utilizing high-performance detectors can generate spectra (4000–200 cm^{-1}) in 60 sec. Sample amounts of 10 μg were found to be sufficient for such spectra.

B. Fourier Transform Infrared Spectroscopy

Fourier Transform IR spectroscopy is widely used because of its rapidity of providing high-resolution spectra with samples in the nanogram range. A complete spectrum for a sample of 10^{-9} g can be obtained in less than 1 sec.

FTIR permits rapid quantitative characterization of solids, liquids, and gases.

Kinetic processes can be monitored by a technique known as time-resolved spectroscopy which involves FTIR. This method has been applied to analysis of complex materials such as polymer film stretching which can be carried out in milliseconds and chemical transformations involving, for example, coal pyrolysis; it also permits on line analysis of products subject to chromatographic separation methods such as GC and LC. During the past five years GC–IR and GC–FTIR involved separating of mixtures and analysis of the individual compounds by IR spectroscopy. The sensitivity limitation of IR detectors with respect to GC and the time difference between the elution of a GC peak (measured in seconds) and the time scan were two of the problems encountered. GC–FTIR allows an IR spectrum taken from a 5-μg GC peak of isobutylmethacrylate by repeatedly scanning with spectral accumulation and enhancement (Fig. 8). FTIR measurements may be carried out by one of the following techniques: (a) KBr pellets, (b) photoacoustic, and (c) diffuse reflectance methods.

The rapid-scanning property of FTIR spectrometers is having its greatest impact in the field of GC–FTIR. FTIR is now rivaling mass spectrometry for the identification of unknowns eluting from gas chromatographs and in one study was shown to identify more peaks than MS. It is clear that GC–FTIR will be of importance for the analysis of environmental samples. The interface between FTIR and HPLC is still at a premature stage, and most of the HPLC–FTIR results reported to date have involved the use of size-exclusion chromatography with chlorinated solvents, since these have good transmission over much of the infrared. The maximum concentration of most peaks eluting from either normal-phase or reverse-phase HPLC columns rarely exceeds 10 ppm, so that on-line detection in a flow cell is very difficult. For normal-phase and reverse-phase HPLC, continuous elimination of the solvent is usually required before identifiable IR spectra can be measured; no commercial instruments are yet available. For at least one of these systems, a detection limit of 100 ng has been reported for reverse-phase separations with a water–methanol mobile phase.

In conventional HPLC–FTIR, the interferograms are continuously recorded and stored during the analysis. Next, the absorption bands due to the solvent are subtracted from the solution spectra. In order to facilitate these measurements, the volume of the flow cell must be of the order of 1 μL, implying that even with the packings of 3–5 μm diameter, the evolution volumes will be significantly higher than the cell volume. Thus, only a small fraction (less than 1%) of each separated component will be in the cell when the measurements are made. Of the several interfaces proposed for solvent elimination, the most promising one involves an initial concentration step in the concentrator tube using N_2 gas and above ambient temperature, followed by the deposition of the concentrated solution on the KCl powder. Further elimination of the solvent is achieved with a stream of air.

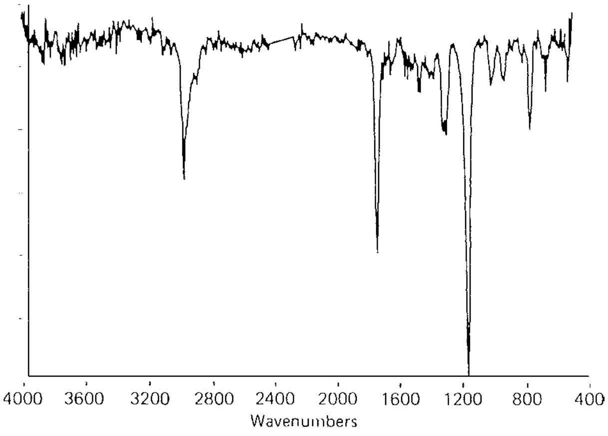


FIGURE 8 IR spectrum of a small GC peak (5 μg) of isobutylmethacrylate using spectral accumulation (FTIR/GC cell).

1. Diffuse Reflectance Fourier Transform Spectrometry (DRIFTS)

Diffuse reflectance Fourier transform spectrometry (DRIFTS) is a multifaceted technique for studying solid samples. Diffuse reflectance has been known for some time as a solid sampling method. When infrared radiation falls onto a sample surface, one of the following processes can occur. (a) The radiation can be absorbed; (b) it can be reflected from the sample (specular reflectance); or (c) it can penetrate the sample before being scattered (Fig. 9). This latter effect is known as diffuse reflectance, and it is this radiation that the technique of DRIFTS measures. The technique is applicable to most solids, involves very little sample preparation, and avoids the unnecessary complications of absorptions due to solvents or mulling agents. Its greatest advantage by far, however, is its ability to handle a wide range of samples, including conventional

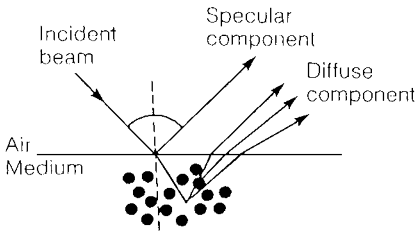


FIGURE 9 The difference between specular reflectance and diffuse reflectance from a solid surface.

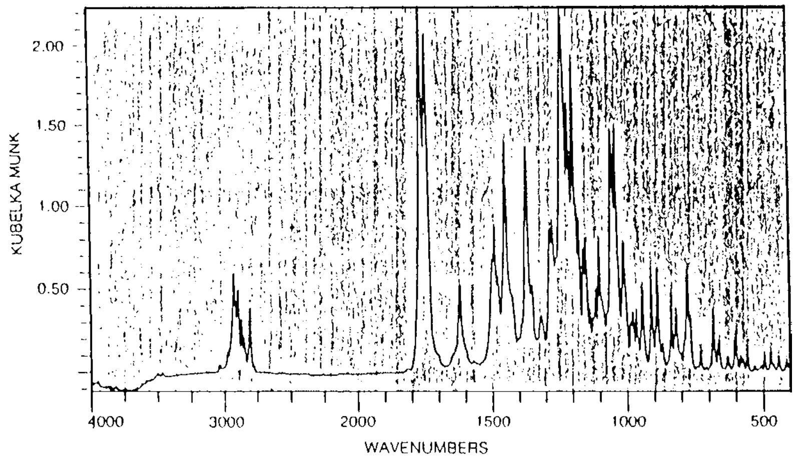


FIGURE 10 Infrared spectrum of heroin by diffuse reflectance spectroscopy as a mixture in potassium bromide using approximately 1 mg of heroin.

powders and crystals, solids with rough surfaces, gemstones, minerals, plastics, and fibers. The best results are achieved with the sample in powdered form. The quality of the spectra can be enormously enhanced by mixing the sample with potassium bromide in the ratio of 1:10 for organic materials.

Many pharmaceuticals can be studied successfully using this technique, for example 1 mg of heroin in potassium bromide (Fig. 10). It is possible to monitor the concentration of heroin in a potassium bromide sample once a calibration graph has been prepared. The DRIFTS technique can be applied to quantitative analysis.

Another application of DRIFTS is the identification of fibers. There are many types of fibers, both synthetic and natural, which can be identified by IR spectroscopy but not so easily by other techniques.

By using a micro sampling cup it is possible to produce diffuse reflectance spectra with a few milligrams of the sample, such as less than 1 mg of heroin. The DRIFTS technique can thus be applied to the screening of illicit street drugs, such as a mixture of cocaine and amphetamine.

2. Two-Dimensional Infrared Spectroscopy (2D IR)

Two-dimensional infrared (2D IR) spectroscopy is a recent novel technique based on time-resolved IR spectroscopy. The spreading out of IR spectra over two dimensions simplifies complex spectra with overlapping peaks. Correlation of such peaks may identify or eliminate specific intermolecular and intramolecular interactions. Such correlations can assist in assigning peaks to particular vibrational modes. In 2D IR, a system is excited by an external perturbation, which induces a dynamic fluctuation of the IR spectrum. A correlation analysis is applied to the time-dependent IR signals to yield a spectrum defined by two independent wave numbers. By spreading IR peaks over the second dimension, a complex spectrum consisting of overlapped peaks can be substantially simplified, and spectral resolution is enhanced. Peaks located on a 2D spectral plane provide information on connectivity and interactions among functional groups associated with the IR bands. The technique was recently developed by Isao Noda and his co-workers, and he used it to demonstrate the known incompatibility of polyethylene and polystyrene by showing that the two resins are segregated at the molecular level (Fig. 11).

3. Chromatographic Optimization of Combined GC–FTIR–MS

Recent rapid developments of analytical instrumentation have made it possible to obtain simultaneously GC–MS and GC–FTIR information on a sample from one injection and one GC separation. Combining GC–MS and GC–FTIR into one system provides a higher confidence qualitative result and saves time, laboratory space, and money. It should be realized that mass spectrometry and infrared spectrometry are complementary.

In terms of relative practical sensitivity, the Mass Spectrum Detector (MSD) in the scanning mode is approximately 10 times better than the Infrared Detector (IRD). Thus, a mass spectrum on a level of 1 ng requires 10 ng for an IR spectrum if the compound is a strong absorber.

4. Matrix Isolation Fourier Transform Spectrometry (MI-FTIR)

Matrix isolation techniques are used in the spectroscopic detection of (a) chemical species that are too reactive, too unstable, or too transient to be investigated under

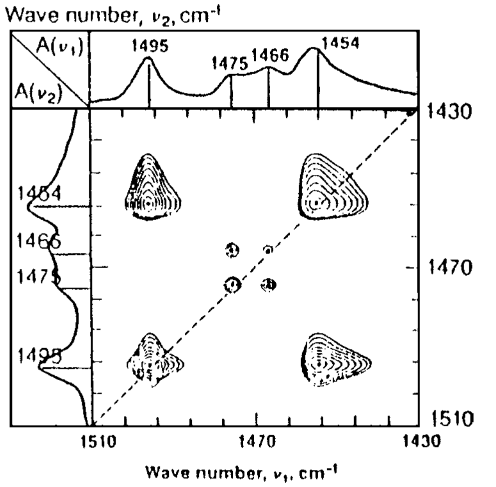


FIGURE 11 Contour map of 2-D IR spectrum shows incompatibility of resins. Margins show infrared spectra in usual form. Auto-peaks appear along diagonal at identical coordinates because each peak is the same as itself. Peaks at 1454 and 1495 cm^{-1} come from polystyrene. Peaks at 1466 and 1475 cm^{-1} come from polyethylene. Square pattern of off-diagonal cross-peaks with polystyrene auto-peaks shows that relaxation rates of polystyrene groups are correlated with one another through interaction. Square pattern of off-diagonal cross-peaks with polyethylene auto-peaks shows that relaxation rates of polyethylene groups are correlated with one another through interaction. Lack of off-diagonal cross-peaks of polystyrene in square patterns with auto-peaks of polyethylene (and vice versa) shows lack of correlation of relaxation times and therefore lack of interactions among polystyrene and polyethylene groups.

laboratory conditions and (b) identification of unknown components in complex mixtures.

Although MI techniques have been used for more than 30 years, it is during the past decade, with the advent of FTIR spectrometers, that MI-FTIR has been used to detect organic transient species such as free radicals, hydrogen bonding, complexes, and reactive intermediates. Matrix isolation is a route for embedding a sample in a low-temperature matrix. In this method, a sample vapor stream is codeposited onto a cryogenic surface with a diluent inert gas such as argon or nitrogen. Dilution factors of 10^4 or more eliminate solute aggregation, and phonon-coupling is not a serious problem with the common matrix gases. For example, the less stable of the two conformers of 1,3-butadiene can be observed by depositing the same in an argon matrix at 10 K directly from an oven held at 1200 K, thus trapping the high-temperature equilibrium mixture. Analyses of isotopically labeled gas mixtures demonstrated that MI-FTIR was superior to gas chromatographic methods for the qualitative analysis of isotopic molecules. Excellent selectivity employing MI-FTIR was observed by the spectroscopic separation of the three isomers of ethylene-d_2.

a. Analytical applications using GC–MI-FTIR. If GC–MI-FTIR spectrometry is to become a useful analytical tool for the identification of unknown components in complex organic mixtures, matrix effects on the appearance of FTIR spectra must be well characterized. The requirements of using GC–MI-FTIR for analytical analyses are:

1. The spectra of individual compounds must have narrow bandwidths and high resolution.
2. The FTIR spectrum of each compound must be unique, conclusive, and relatively simple.
3. The spectral features for a given compound must be reproducible from sample to sample.
4. The molar absorptivity of the compound must be independent of concentration and not be affected significantly by other sample constituents.
5. The sampling medium must be transparent, inert, and chemically pure.
6. The sample must be soluble in the medium.

The disadvantages of GC–MI-FTIR involve high initial cost, it is primarily limited to volatile samples because of the difficulty in eliminating the solvent, and it is a two-step process as opposed to on-the-fly spectral aquisition, causing an increase in the time required for each analysis.

GC–MI-FTIR spectrometry has also been used in the detection of polychlorinated dibenzo-*p*-dioxins (PCDDs), but because of the presence of 76 PCDD isomers, which include 22 isomers of the tetrachlorodibenzo-*p*-dioxins (TCDDs), such analyses are not easy. In 1986 it was shown that the sensitivity of GC–MI-FTIR for TCDDs in the picogram range were comparable to those obtained by GC–MS for the determination, for example, of 2,3,7,8-TCDD in an environmental sample. The spectrum of the isomer 1,2,3,4-TCDD was clearly identified at a level of 0.156 ng but required the coaddition of 5000 scans. On the other hand, an identifiable spectrum of ethyl carbamate in whiskey extracts was obtained with only 300 scans for 0.6 ng of the material. The levels of ethyl carbamate as determined by GC–MI-FTIR were in good agreement with those obtained from GC–MS–MS.

Recent developments in expanding MI techniques include (a) the combination of gas chromatography and MI-fluoresence spectrometry, and (b) the combination of supercritical fluid chromatography and MI-FTIR.

b. Protein sample preparation. A device for isolating material separated by liquid chromatography to be subsequently analyzed by FTIR spectroscopy works as follows. The eluent from the chromatograph is sprayed

from a heated, linear capillary nozzle toward the surface of a slowly rotating germanium disc. The microdrops are constrained by a heated gas sheath that also rapidly vaporizes the solvent, leaving a particle stream of solute material that collects on the disc surface. Instrumentation based on this technology is used for a wide variety of applications in polymer, pharmaceutical, and environmental laboratories.

The recent development of electrospray ionization (ESI) and matrix-assisted laser desorption-ionization (MALDI) sources for creation of macromolecular ions has brought about a renaissance in the field of mass spectrometry (MS).

It is now common to determine high molecular weights such as proteins to within a few Daltons.

MS-based methods for sequencing are proliferating and replacing or complementing traditional biochemical methods for many applications. Mounting evidence suggests that some conformational properties of biomolecules in solution are preserved during the ionization process and persist over the transience time that ions exist in mass spectrometers.

In the past 5 years a new field involving structural studies of biomolecular ions has emerged. Studies of biomolecular ions in the gas phase provide information about the nature of conformation in the absence of solvent:

The structural studies of protein ions have been grouped into two categories: chemical probes, in which structural characteristics are deduced by monitoring the products of ion-molecule reactions, and physical probes such as measurements of dissociation patterns and measurements of cross sections.

An important factor in folding of proteins involves solvent–molecule interactions. Proteins appear to form physiologically relevant conformations in the presence of a minimum hydration shell.

Recent studies have considered structural changes that occur when proteins are lyophilized by examining anhydrous powder and thin films. Techniques such as FTIR spectroscopy, Raman spectroscopy, and H-D exchange NMR methods have shown structural changes upon removal of solvent.

When the proteins are heated, they denature, making them less volatile; the molecules decompose rather than vaporize. ESI has offered a simple, efficient means of producing gaseous, anhydrous biomolecules for study by MS. Thus, anhydrous proteins also fold and unfold, and new conformations are established in a grossly simplified environment whose only intramolecular constraints are relevant.

Charged droplets containing molecular ions are initially formed by pumping a solution through a narrow capillary tube that is electrically biased with respect to an entrance orifice of a MS. The electrode shapes, geometries, and potentials are important for penetrating a stable spray of droplets; however, many different experimental configurations can be used. Two widely held views of the mechanism for ESI are that solvent-free molecular ions are formed either after a series of droplet fissions, followed by solvent evaporation from the ion, or by ion evaporation from the droplet surface.

C. Fourier Transform Raman Spectroscopy (FT-RS)

Raman spectroscopy (RS) has been known for more than 60 years. It is during the past decade that the organic chemist has included RS as an additional tool for structure determination. The Raman spectrum provides complementary information to the IR spectrum in that both are vibrational spectra. Often the weakest signals in the IR spectrum are the strongest signals in the Raman spectra. Modern Raman spectrometers operate with a monochromatic coherent laser light source. Sampling handling is much easier in RS than in IR spectroscopy. The sample can be dissolved in a number of solvents since their Raman absorptions are more limited than in the IR. Solid samples can also be examined as powders or crystals. Liquids can be placed in small quartz or Pyrex glass capillaries, and the scattered radiation is observed through the open end of the capillary. Samples can also be examined without removing them from their bottles. If several spectroscopic analyses are to be performed on a small amount of an organic sample, then the RS is obtained first because it is totally nondestructive and does not require dilution in solvents, as in the case of UV or NMR. Certain group frequencies such as primary amines, alkynes, and nitriles are strongly absorbed in the RS but are difficult to observe in the IR spectrum. Raman spectroscopy has a number of experimental advantages over IR: (a) Window problems hardly exist, if visible or near IR lasers are used as sources; glass is a robust window material with excellent transmission; (b) since transmission through the sample is not involved, very little preparation is required in RS; (c) water, which has an extremely strong IR spectrum, produces only a weak RS and is the solvent of choice in biochemical and aqueous studies.

Raman spectra are often plagued by the problem of fluorescence when conventional shorter (visible) wavelength is used. It was estimated that from 1966 until 1989 less than 20% of the samples studied gave suitable spectra with visible laser Raman dispersive spectrometers. Even so, considerable time is wasted in obtaining a reasonable Raman spectrum. Since 1987 Near Infrared Fourier Transform Raman Spectroscopy (NIR-FTRS) has been

available. This technique has the potential to be more versatile than conventional laser Raman spectroscopy. The technique in NIR-FTRS is to excite the spectrum with near IR laser, pass the scattered light through filters to remove the reflected and Rayleigh scattered excitation radiation, and then to process the Raman scatter on an FTIR instrument. Being infrared excited, the spectra were essentially clear of fluorescence, while color due to absorption was far less of a problem in near infrared than with the conventional techniques. For example, rubrene, which is deeply colored and intensely fluorescent, produced a well-defined Raman spectrum at 1064 nm. Recently FTRS spectra of alkaloids such as heroin, codeine, and morphine have been obtained, but because of the complexities of their spectra they have not yet been fully analyzed. It is obvious that within the next few years group frequencies characteristic of structural features in complex organic and bioorganic compounds will be determined. It appears that NIR-FTRS is the choice method in obtaining spectra of biochemical substances such as enzymes, proteins, and polypeptides because such molecules can be studied in aqueous media. Interpretation of the complex nature of these biomolecular spectra will also take time. Commercial instruments offering mid-infrared (400–4000 cm^{-1}), near infrared (4000–10,000 cm^{-1}), and Raman spectra all on the same spectrometer will be available in the near future.

D. Vacuumless Spectrochemistry in the Vacuum Ultraviolet

The emission wavelength region below 200 nm has traditionally been called the vacuum ultraviolet (VUV). This region contains the best spectral lines for many industrially important elements, such as carbon, phosphorus, sulfur, and boron, in steel and cast iron by spark excitation, and the halogens for the inductively coupled plasma (ICP) analysis of oils and fuels.

Wavelengths below 200 nm are absorbed by some components of air, primarily oxygen and water vapor. Therefore, the usual approach to spectrochemistry in this wavelength region is to evacuate the optical system.

The use of a nitrogen gas-filled system provides important advantages for both the industrial and academic spectroscopist. This technology opens up the low wavelengths of the emission spectrum which contains many important spectral lines for elements of interest for practical spectrochemistry. It should be emphasized that none of the wavelengths below 140 run have been achieved with a conventional vacuum spectrometer.

One of the economic advantages of ICP analysis of halogens in oils, fuels, and other matrices is that it is faster than other methods such as X-ray fluorescence (XRF).

As technology advances, it provides new opportunities. The gas-filled optical system technology opens up the VUV for spectrochiral analysis. The terminology "far ultraviolet" might be more appropriate since it includes access to this wavelength region by gas-filled optical techniques.

E. Mass Spectrometry (MS)

Mass spectrometry provides a means for studying samples at the molecular level. Although MS is basically a structure-identifying tool, it is not as specific for the detection and determination of functional groups (IR) or relative positions protons or ^{13}C nuclei (NMR) spectroscopy. Mass spectrometry, like UV, IR, and NMR spectroscopy, works best with pure samples. Structural elucidation of mixtures is not possible without some means of separating each constituent in the mixture as in GC–MS.

Besides being a useful structure-elucidating tool, MS can be applied to detect very low levels of specific compounds and elements. Accurate determination of masses can also be determined. An important advantage of the MS technique is its high sensitivity and accuracy. MS is able to provide more specific information per given amount of material than any other analytical technique. Furthermore, even with minute amounts of material (10^{-6}–10^{-9} g), this information can be provided in a reproducible and accurate manner. In the 1940s MS was an important analytical technique for characterizing complex fuel mixtures. About 25 years ago interest was centered on a systematic study of ionic fragmentation mechanisms in order to provide a set of rules that could be applied to elucidate structures of organic compounds. In the late 1960s mass spectrometers were coupled to gas chromatographs. This provided such dramatic improvements in selectivity and sensitivity that MS has since become one of the most generally useful analytical techniques for identification and quantitation of organic substances at ultratrace levels. Chemical ionization mass spectrometry (CIMS) uses reagent ions rather than electrons to ionize a sample. Field desorption and rapid heating techniques are extending mass spectrometry to high-molecular-weight, low-volatility compounds. During the past decade mass spectrometry has undergone a number of significant changes that have far-reaching contributions especially in biology and medicine. The mass range of mass spectrometers has been extended by approximately an order of magnitude in the past decade. Certain types of mass analyzers have been used to reach higher masses (150,000 in some quadrupole experiments). Commercial instruments are now available with mass ranges of 7500 compared with 1000 about ten years ago. Another improvement is desorption ionization [including fast atom bombardment

(FAB), secondary ion mass spectrometry (SIMS) and fission fragment methods] which allows ionic, nonvolatile compounds to be examined by MS. A further development is the integration of separation and analysis techniques represented by tandem mass spectrometry (MS–MS) and the much improved liquid chromatography-mass spectrometry (LC–MS) interfaces. The latter capability is already proving its effect with respect to GC–MS but without the limitation to volatile, low-mass compounds. MS–MS is a two-dimensional form of spectrometry which often improves signal-to-noise ratios as well as providing entirely new capabilities, such as that of scanning a mixture for all constituents having particular structural subunits.

The GC–MS combination has had considerable impact on biological and environmental research, providing a specific means of characterizing constituents of mixtures and having sufficient sensitivity and quantitative accuracy to trace constituents. The recent discovery in 1977 of the neural excitotoxin, quinolinic acid, in the mammalian brain at the level of nanomoles per gram was achieved by GC–MS of the volatile hexafluoroisopropanol diester derivative with electron impact (EI) ionization. Quantitation was based on standard addition and single-ion monitoring. The GC–MS technique was applied in the discovery of 19-hydroxylated E prostaglandins by first protecting the unstable β-ketol system by oximation. This discovery, in 1974, implied that the previously identified prostaglandins could be artifacts.

Challenging structural problems. including protein structure determinations, can often be solved by a combination of techniques. Two mass spectrometric methods were used to deduce the structure of the 112-amino-acid antitumor protein macromycin derived from streptomyces culture. Partial acid digestion gave a mixture of di- to hexapeptides, which was derivatized and analyzed by GC–MS with EI.

Tandem mass spectrometry has been applied in the discovery and confirmation of the structure of the metabolites of the drugs primidone, cinromide, and phenytoin in plasma and urine extracts; analyses were completed in less than an hour, using concentrations of 1 to 50 μg/mL. The study is predicted on the speed and flexibility of MS–MS scans made with a triple quadrupole instrument, and on the realization that metabolites often retain a large portion of the parent drug structure.

Analysis of complex mixtures has often involved time-consuming procedures such as extraction, centrifugation, and chromatography. MS–MS and LC–MS are recent developments that reduce such time-consuming methods. By linking two mass spectrometers in tandem it is possible to employ the first as a separator and the second as an analyzer and, hence, to perform direct analysis of mixtures. The two principal advantages of this system can be illustrated by considering a complex coal liquid mixture. The signal due to a dioxin spike is lost in the chemical noise from the other constituents and single-stage mass spectrometry is not capable of analyzing for it. MS–MS filters against chemical noise and allows a high-quality spectrum of the dioxin to be recorded. In addition to improving detection limits in this way, tandem mass spectrometry provides alternative scan modes which can be employed to search the data bank for particular information. For example, chlorinated dioxins are characterized by the loss of COCI, so a scan for the reaction reveals all dioxins present in the mixture. The GC–MS–MS system can decrease GC–MS detection limits by an order of magnitude. It has been reported that the drug, isosorbide-5-mononitrate, a coronary vasodilator, is metabolised to the glucuronide, which can be determined in urine by a simple MS–MS procedure to 0.1 ng/mL. The improved detection limit in the MS–MS experiment is the direct result of minimizing interferences. Both the sensitivity and the speed of analysis with MS–MS can be illustrated with the administration of the drug tetrahydrocannabinol in doses of 0.1 mg/kg for eight days down to 10^{-11} g/ml by using a combination of GC–MS with simple MS–MS to avoid extensive sample cleanup. Sensitivities in parts per trillion have been reported in MS–MS studies on animal tissue. High-resolution mass spectrometry and MS–MS have been used to achieve absolute detection limits of less than 10^{-12} g for tetrachlorodibenzodioxin and a GC–MS–MS combination has produced spectra with 250 fg samples ($<$20 parts per trillion) at the rate of 30 samples per day. In terms of sample throughput, the determination of trichlorophenol in serum at concentrations as low as 1 ppb and a rate of 90 samples per hour was reported. Characterization of the foodstuff contamiant (and chemical warfare analog) vomitoxin at 25 pg in wheat at a rate of 10 minutes per samples was also reported.

A different approach to the characterization of mixtures of nonvolatile compounds is LC–MS. The first practical LC–MS interface was based on complete removal of the solvent and temporary storage of the solute during transport by a moving belt or wire into the ion source. In the source the sample is either thermally desorbed and ionized by electron or chemical ionization or the belt is bombarded by an energetic beam to create secondary ions. The large pumping capacities of chemical and atmospheric ionization sources make it possible to work at flow rates consistent with normal column operation, for example, 2 mL of aqueous mobile phase per minute. The solvent itself acts as the reagent gas in these experiments. An alternative, the thermospray procedure does not use any externalionization technique. An aerosol generated in the interface is

evaporated, and separation of charges present in the nominally neutral solution allows positive- and negative-ion mass spectra to be recorded. The performance of LC/MS for β-hydroethyltheophylline is 10 pg (selected ion monitoring) or 1 ng (full spectrum) with respect to its detection limits. The method involving direct liquid introduction and CI gives comparable data; for example, 50 ng of vitamin B_{12} gives a negative-ion spectrum of high quality.

Fourier transform mass spectrometry (FTMS) illustrates the speed with which instrumental developments are transforming mass spectrometry. The high resolution of FTMS is probably its most important aspect. Impressive performance data have been reported, such as a resolution of 1.4×10^6 for $m/z = 166$ from tetrachloroethane and 10^8 for $m/z = 18$ from water. FTMS instruments are capable of performing MS–MS experiments. Unlike a conventional MS–MS experiment, in which the different stages of analysis are separated in space, the separation here is achieved in time. This allows the extension of the experiment to three (MS–MS–MS) or more stages. Because FTMS instruments require very low pressures for optimum performance, interfacing with chromatography is a problem, although a GC–FTMS has been reported by Wilkins in 1982.

One aspect of LC–MS which is rapidly being developed is the HPLC–MS and has great promise for the analysis of many compounds. The primary obstacle to easy coupling of HPLC to MS arises from the fact that the flow for conventional HPLC columns is approximately an order of magnitude greater than can be accommodated by the commonly used ion sources and pumping systems. The requirement for a transfer of a maximal quantity of sample and a minimal amount of solvent into the ion source of the mass spectrometer has necessitated the development of different interfaces. The need for introduction of smaller amounts of HPLC effluents into a mass spectrometer has stimulated the development of narrow bore columns. The application of HPLC–MS with a direct liquid introduction has been applied to the identification of marine sterol peroxides by Djerassi and Sugnaux in 1982. An ultrasphere ODS column of i.d. 5 μm and methanol–water (99:1, v/v) solvent system were used.

At present it is difficult to predict if the HPLC–MS technique with still many difficulties will become a routine tool for detection and identification of HPLC solutes. It is clear, however, that this technique remains the only coupled MS method by providing unambiguous identification of solutes in HPLC effluents at room temperature or even lower than room temperature. Many natural products that have been identified by GC–MS techniques can be verified by HPLC–MS.

F. GC–MS–MS

Full-scan and selected ion monitoring (SIM) represent at least 95% of the MS analysis performed today. Full-scan analysis has always offered three benefits: (1) selectivity based on m/z detection, (2) universal detection across a wide operating range and (3) library searches (spectral comparison) for confirmation of identity. Since the source of the data for SIM is the same as for full scan, SIM offers no solutions to interference problems. The primary advantage of SIM has always been its improved signal-to-noise ratio (S/N) versus full scan and its interpretation simplicity (ion ratios).

When MS–MS is operated in a selected reaction monitoring (SRM) mode, the first step is identical to SIM. As a result, isolation of the parent (precursor) ion offers no better data than SIM, assuming equivalent spectral resolution. If these were interferences in the full scan and SIM data, these would also be observed as increased intensity of the parent ion for MS–MS.

The second step, collision-induced dissociation (CID), does offer a solution to this problem. The amount of fragmentation in CID varies with the collision energy, but under all conditions, the secondary ions (daughters or product ones) have a direct, unique relationship to the parent ion.

Quantitation based on the MS–MS daughter ion yields accurate results, even in the presence of an interference. Qualitative decisions are possible in MS–MS based on library searches against user libraries or ion ratioing techniques similar to SIM. MS–MS is actually a more powerful tool for structural elucidation of unknowns that do not search well against standard libraries.

Detailed descriptions of the tandem-in-time principles of ion-trap MS–MS have been well documented. The critical elements of ion-trap MS–MS and the resultant advantages can be summarized as:

- Steps of the MS–MS process are executed as a timed sequence in a single analyzer, rather than across the path of multiple quadruple analyzer.
- Low cost of only one analyzer; high sensitivity; changes from MS to MS–MS or MS–MS–MS can be fully automated; no special expensive collision gases such as argon; good fragmentation efficiency, even for stable ions, which extends the range of applications.
- MS–MS offers clear, simple, error-free results even with complex samples.

When one realizes that the low picogram and even femtogram detection limits (amounts on column) are common, MS–MS becomes even more appealing.

G. MALDI–MS

Matrix-assisted laser desorption-ionization mass spectrometry (MALDI–MS) is a practically effective tool for the measurement of natural and synthetic polymers. The ability to ionize a broad range of materials, high sensitivity, broad mass range, fast sample preparation, and the absence of fragmentation are principal characteristics of the technique.

The MALDI–MS instrument, directs short duration pulses of laser light at a prepared mixture of a sample and a matrix chemical. The laser energy volatilizes a portion of the matrix/polymer sample (as individual polymer molecules) which are then ionized. These individual ions are accelerated to a fixed energy in an electrostatic field and directed into a field-free flight tube. The ions then impact into an ion detector, and the time intervals between the pulse and laser and impact on the detector are measured.

MALDI–MS has found applications in the mass determination of molecules ranging from small peptides, oligosaccharides, and oligonucleotides, to large proteins and synthetic polymers. Polymeric constituents in coal ranging from several hundred molecular weight to 16,000 have been mass analyzed by MALDI, and synthetic polymers of MW up to 70,000 have been reported.

H. Nuclear Magnetic Resonance Spectroscopy

It was over 35 years ago that NMR spectroscopy became attractive to organic chemists when Knight reported in 1949 that the precise frequency of energy absorption depends on the chemical environment of hydrogens. The frequency of radiation required for NMR absorption depends on the isotope and its chemical environment: the number of absorption peaks for magnetic nuclei and the intensity of the absorption peaks is proportional to the number of nuclei. Before the advent of Fourier transform, NMR spectra were measured by continuously varying the field-frequency ratio seen by the analytical sample and is referred to as the continuous-wave or slow passage method. During the past 15 years FTNMR spectrometers have appeared which have the advantage over continuous-wave NMR in that an increased sensitivity for an equal amount of instrument time is obtained, because the entire spectrum is observed with each pulse.

NMR is a valuable technique in molecular elucidation and verification. In some instances, the NMR spectrum is sufficient for identifying relatively simple unknown compounds while in other applications structural information from NMR spectra compliments that of other chemical and spectroscopic methods. FTNMR is very useful for studying reaction kinetics and chemical equilibria as well as quantitatively analyzing mixtures. For example, it was found that in the reduction of 5,5-dimethylcyclohexen-2-one with $LiAlH_4$ two products were obtained, 5,5-dimethylcyclohexen-2-ol and 15% of 5,5-dimethylcyclohexanol. This impurity was not previously detected by other methods.

It was only from FTNMR that chemical kinetics and structural analyses of the acid-catalyzed rearrangement of *trans*-methyl chrysanthemate to methyl lavandulyl esters could be accurately analyzed.

Simplification of NMR spectra of more complex molecules such as natural products and polymers can be done in three ways:

1. Selective decoupling spin–spin interactions by the double resonance technique. The sample is simultaneously irradiated with two radio-frequency fields; one frequency gives the resonance pattern of interest, and the other frequency is the resonance frequency of the coupled nuclei. When the second proton is irradiated with sufficient intensity, the spin–spin coupling collapses giving simple spectra.

2. Deuterium exchange of protons. Substitution of deuterium for hydrogen often gives spectra having no observable coupling to deuterium because of its small magnetogyric ratio. Protons attached to heteroatoms are easily exchanged by deuterium such as OH in alcohols, SH in mercaptans, and NH in amines. The use of reagents containing deuterium such as $LiAlD_4$, D_2SO_4, $NaBD_4$, CD_3COCl, D_2O and $CD_3ND_2 \cdot DCl$ are just a number of reagents that can be used in synthesizing complex structures in which the NMR spectra would be much simpler. It should be realized that selective decoupling can then be applied to deuterated compounds, thus producing even simpler spectra. Simple proteins can be examined in this way.

3. Complexation of the organic molecule by certain paramagnetic metal ions can result in useful changes in the NMR spectrum. It has been shown that certain transition elements such as europium and praesodymium induce large shifts in the proton NMR spectra of organic compounds that coordinate to the metal, for example, molecules that have certain functional groups that contain oxygen or nitrogen. The chloroform soluble *tris* (dipivalomethanato)europium (III) complex, $Eu(DPM)_3$, has been used as a shift reagent to simplify the proton NMR spectra of compounds, including alcohols, steroids and terpenoids such as the racemic mixture of the *cis*-δ-lactone of cis chrysanthemic acid (Fig. 12). In this instance the singlets of the methyl moieties are clearly shifted in the presence of the enantiomer (−)-2,2,2-trifluoro-1-(9-anthryl)ethanol or Pirkle's reagent (Fig. 13).

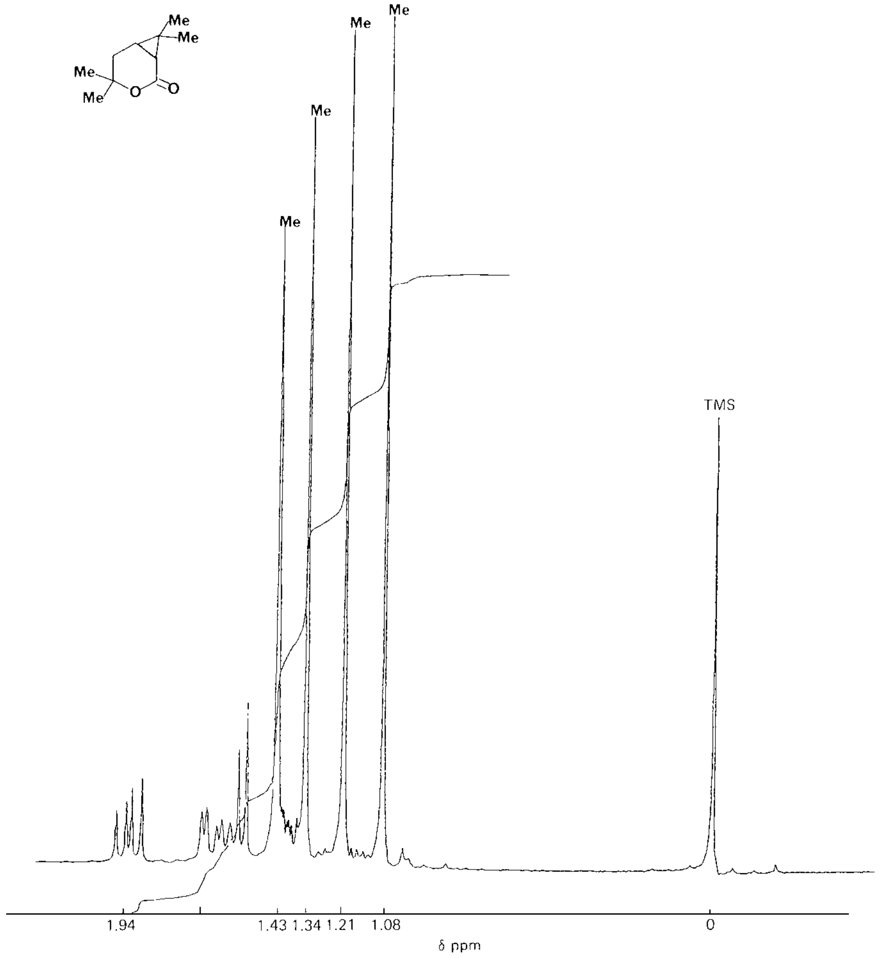


FIGURE 12 ^{1}H NMR spectrum (300 MHz) of (±) *cis*-δ-lactone in $CDCl_3$.

A routine sample on a 60-MHz instrument would require about 5–50 mg of the sample in about 0.4 mL solvent. A microtube consisting of a thick-wall capillary allows spectra to be obtained on less than 1 mg of sample. The use of FTNMR makes it possible to obtain spectra on amounts as small as 5 μg. The ideal solvent should contain no protons in its structure, be low-boiling point, nonpolar, and inert. Carbon tetrachloride is often used in non-FTNMR spectroscopy, depending if the sample is sufficiently soluble in it. The most widely used solvent is deuterated chloroform ($CDCl_3$).

I. Nuclear Magnetic Resonance (NMR)

New health and materials sciences for societal needs will both fuel the drive to make high-field NMR spectroscopy a more accessible research tool. Currently, only a handful of the most powerful 800-MHz NMR spectrometers and no more than a dozen 750-MHz instruments are in use in the United States. More powerful instruments are in the works, including a 900-MHz and gigahertz instruments being built in the United Kingdom, United States, Germany, and Japan.

The new instruments will find cures for diseases, resolve the molecular structures of protein, amorphous polymers, ceramics, and biomaterials, and elucidate the dynamics of biological systems.

High-resolution NMR may serve as an ideal compliment to X-ray crystallography because it provides information on molecular dynamics that crystals, being static, cannot. NMR enables observation of changes in the dynamics of a protein when it binds to DNA, or details the relationship between catalytic function and motion. This information would be useful in drug design.

Stoking the interest in faster access to protein structure are the rapid advances in genome research that are spewing out genes with no known function.

A Swiss NMR technique makes it possible to obtain high-resolution solution NMR spectra of macromolecules and supramolecular structures with masses up to several

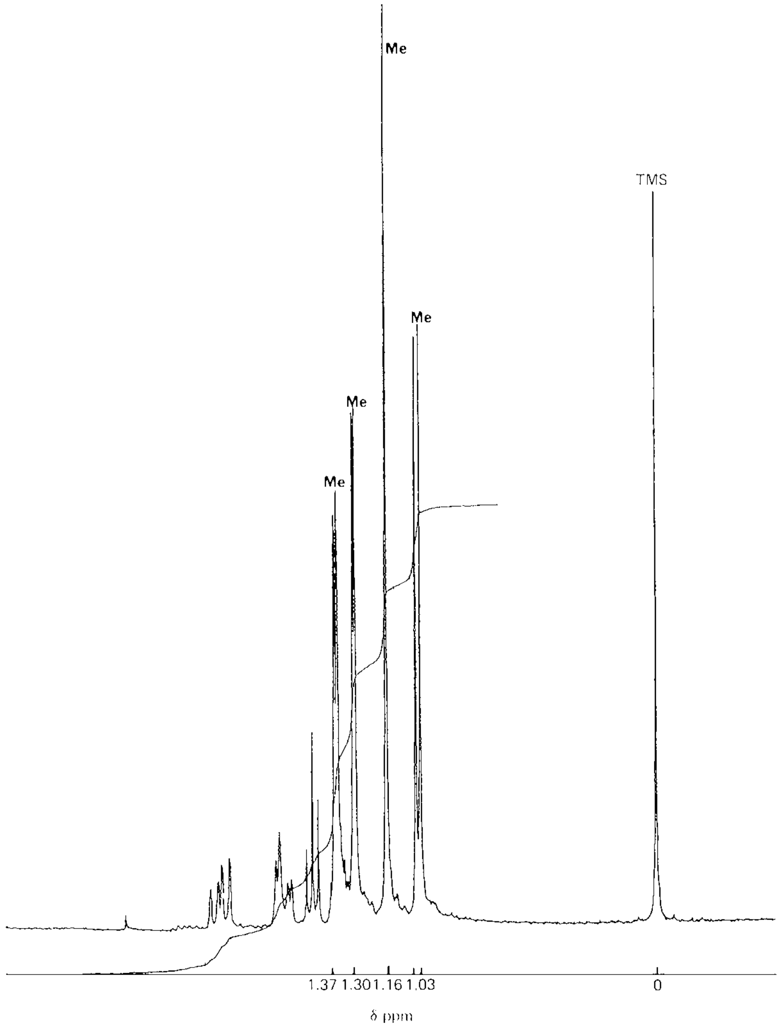


FIGURE 13 ^{1}H NMR spectrum (300 MHz) of (±) *cis*-δ-lactone in presence of (−)-2,2,2-trifluoro-1-(9-anthryl)ethanol (Pirkle's reagent).

hundred thousand Daltons—many times larger than those accessible by conventional NMR techniques.

The technique could prove useful for

- Chemical shift mapping (NMR studies of interactions of protein–protein or protein nucleic acid interfaces in supramolecular assemblies)
- Acquiring structure–activity relationship data on molecular interactions with large receptor molecules (SAR by NMR)
- Obtaining high-resolution NMR spectra of membrane proteins reconstituted in micelles.
- determining NMR solution structures of large macromolecules.

A technique called transverse relaxation-optimized spectroscopy (TROSY) now greatly reduces NMR line broadening with increasing molecular size, permitting NMR analysis of molecules far beyond 100 KDa in size.

1. ^{13}C NMR Spectroscopy

The reason that ^{13}C NMR was developed much later than the proton NMR was due to the low natural abundance of the ^{13}C nucleus (1.108%) as compared to ^{1}H NMR (100%) and the low gyromagnetic ratio. These two problems were overcome during the past 15 years due to the arrival of Fourier transform where the accumulation of scans over a period of time overcomes the problem of the low abundance of ^{13}C. The advantages of ^{13}C NMR compared to ^{1}H NMR is that better resolution is obtained since the ^{13}C absorptions for most of the organic molecules are spread over 200 ppm instead of 10 ppm. Secondly carbons bearing no protons are revealed and finally a count

of the number of protons attached to each carbon results from comparison of the broad-band decoupled ^{13}C NMR spectrum with the off resonance ^{13}C spectrum. Therefore, the number of methyl (quartet), methylene (triplet), methinyl (doublet), and quarternary (singlet) carbons in a fairly complex molecule such as in the natural sweetener, stevioside, $C_{38}H_{60}O_{18}$ (Fig. 14) is easily determined by ^{13}C rather than by ^{1}H NMR.

On the other hand there are a number of disadvantages in ^{13}C NMR. A larger sample size (up to 100 mg) and longer sampling time (up to several days for small samples) is required. If 100 mg of a sample is available and that sample is of high solubility, the time required may be only a few minutes. A good spectrum can be obtained on as little as a 1 mg sample when several days are available for scanning the sample provided the sample is completely stable in the deuterated solvent. Another disadvantage is that due to variations in relaxation times and nuclear Overhauser effects (NOE), the areas of absorption differ for individual carbons (up to a factor of about 10). Thus, it is not as easy to tell relative numbers of carbons from ^{13}C as it is protons from ^{1}H NMR. Finally protons attached to heteroatoms are not visible. For the time being ^{13}C NMR spectroscopy is a method that compliments other spectroscopic techniques necessary to elucidate structures of organic molecules. One commercial application of ^{13}C NMR is the determination of the purity of certain types of organic compounds. For example, whether a reduction, oxidation, or esterification processes have gone to completion can easily and also quite rapidly (if sufficient material is available) be determined.

Recent advances of even more sophisticated and more powerful NMR spectrometers have enabled high resolution ^{13}C NMR spectra where signals as low as 3 ppb can be distinguished at moderate field strength and without resolution enhancement. This is evident in the enantiomeric mixture of the alkane farnasane

$$[CH_3{-}\underset{\displaystyle CH_3}{\underset{|}{C}}H{-}(CH_2)_3{-}\underset{\displaystyle CH_3}{\underset{|}{C}}H{-}(CH_2)_3{-}\underset{\displaystyle CH_3}{\underset{|}{C}}H{-}CH_2CH_3]$$

which gave two clear signals for all the carbon atoms except for one and with a minimum separation of 2.9 ppb.

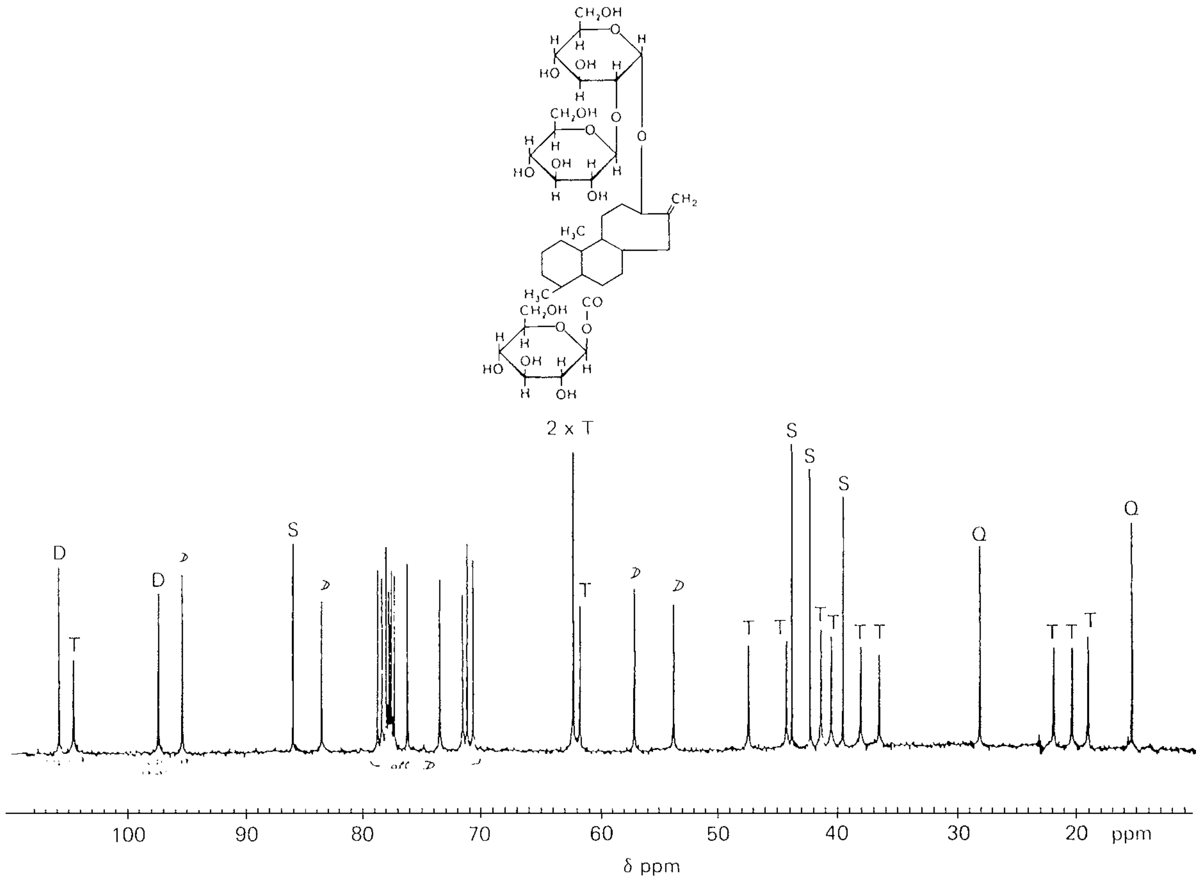


FIGURE 14 ^{13}C NMR spectrum of the natural sweetener stevioside ($C_{38}H_{60}O_{18}$).

Enantiomeric separation and spectral simplification of overlapped resonances in high resolution ^{1}H NMR spectra can be obtained with optically active paramagnetic shift reagents such as $Eu(tfc)_3$. Detection of optical isomers ($\pm$) 2,2-dimethyl-1-phenyl-1-propanol and ($\pm$) carvone, as well as optically active terpenes, can rapidly be determined by the use of optically active shift reagents.

2. ^{13}C-CP/MAS NMR

It has been observed that line broadening of NMR spectra of solid powders arises jointly from ^{13}C–^{1}H static dipolar interactions and from chemical shift anisotropy. These effects are eliminated or greatly reduced by dipolar decoupling (dephasing) of protons and by spinning the sample at an angle of 54.7° (magic angle) to the applied field. For solids a greater rate of spectral accumulation is required, this method is called cross polarization (CP).

High resolution ^{13}C NMR which utilizes both cross polarization with dipolar dephasing and magic angle spinning (MAS) is called ^{13}C-CP/ MAS NMR. This technique directly measures the organic carbon distribution in terms of aliphatic and aromatic carbon structures in oil shales, coals, humic substances, melanoidins, and other polymeric substances.

3. Applications of FTNMR

Because of the known utility of NMR as a fingerprinting technique, there is recent interest in the combination of this method with HPLC. Following the initial work on NMR spectroscopy of static HPLC fractions, considerable efforts have been focused on the development of analyses of flowing liquid chromatographic fractions in the presence of hydrogen containing solvents. The coupling of these two techniques may offer some advantages compared with the HPLC–MS systems since there are no volatility (or molecular weight) requirements, and the HPLC–FTNMR technique is more informative and nondestructive than MS (especially for studies of molecular stereochemistry). Furthermore, both HPLC and NMR measurements are carried out in solution and, thus, no phase transformations are required. Spectra can be obtained at low temperatures for nonstable compounds such as the 19-hydroxy prostaglandins that were previously mentioned. The development of suitable interfaces has been hampered by several problems. Primarily, the intensity of the NMR signal is dependent on the flow rate; the extent of line broadening and, consequently, the spectral resolution, depend on solvent viscosity and detector volume. Because of the relatively low sensitivity of NMR a compromise must be found between the optional flow rate and the detector volume. Moreover, the HPLC–FTNMR operation is limited to certain types of solvents. Since most commonly used HPLC solvents contain hydrogen, the memory of the minicomputer used for data acquisition will be saturated with the solvent signal, thus overshadowing the weak sample signal. This has necessitated the use of special suppression techniques for solvent signals, such as filtering, selective presaturation, selective excitation, and spin-echo techniques, or increasing the dynamic range of the minicomputer. It is possible to use deuterated solvents as the HPLC solvents but this involves considerable expense. During the past decade FTNMR has been applied to solid states especially to elucidate microstructures of polymers. The term microstructure pertains to the level of isomerism of a polymer chain. For example oxidized sites in polymers (carboxylic acids, ketones or alcohols) can be detected and quantified at levels as low as two sites per thousand residues from the ^{13}C NMR spectrum. Although carbon is ubiquitous in synthetic polymers, other nuclei have been used to advantage for polymer structure determination. For example, ^{15}N NMR for nylon polymers, ^{31}P NMR for poly(phosphazenes), and ^{19}F NMR for fluoropolymers. For some time ^{1}H FTNMR has been regarded as less informative for synthetic polymers than the carbon resonance; it is now the focus of renewed interest.

4. Two-Dimensional NMR Spectroscopy (2D-NMR)

The term 2D-NMR is a misnomer because the NMR spectra that have been previously described are two dimensional in the sense that they are plots of signal intensity versus frequency (or its Fourier equivalent, signal intensity against time). By contrast, 2D-NMR refers to spectral data (usually several hundred) that are collected as a function of *two* time scales, t_1 (evolution) and t_2 (detection). 2D-NMR is, therefore, three dimensional: signal intensity versus *two* frequencies F_1 (the Fourier transform in the t_1 time domain) and F_2 (the Fourier transform in the t_2 time domain). The resulting data are then Fourier transformed in both the t_1 and t_2 time domains. The resulting 2D-NMR spectrum is in fact a map containing two frequency axes, the F_1 axis which contains chemical shifts and the F_2 axis which may contain chemical shifts or coupling information or even both. The 2D-NMR spectrum can be displayed either as a *stacked plot* (Fig. 15), which portrays all three dimensions in a perspective view, or a *contour plot* (Fig. 16), which represents a horizontal cross-section of the stacked plot at a selected value of signal intensity.

Heteronuclear shift correlation (HETCOR) is a 2D-NMR technique where a two-dimensional map is

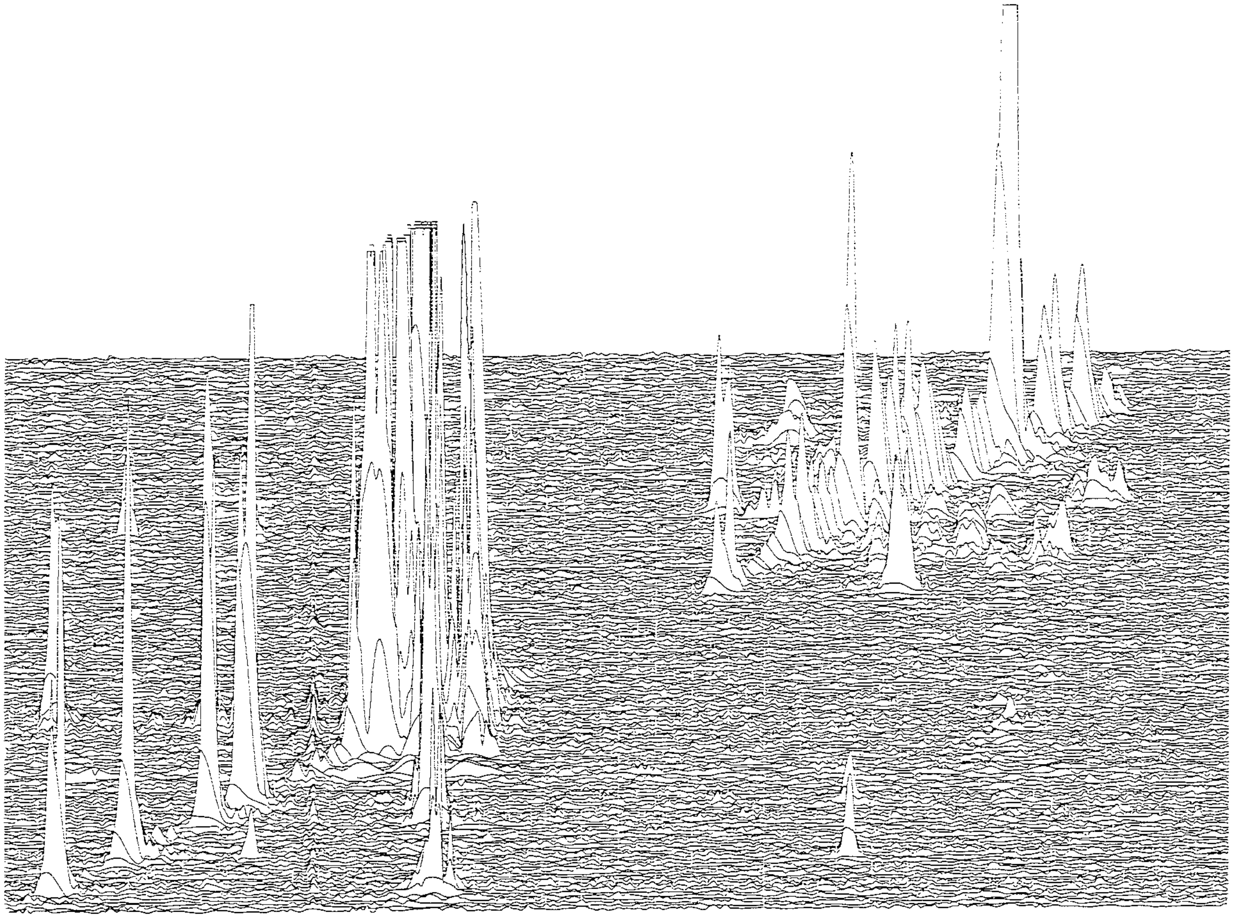


FIGURE 15 COSY NMR spectrum of stevioside (stacked plot).

generated in which the F_2 axis corresponds to ^{13}C chemical shifts (δ_c) or other nonproton nuclei and the F_1, axis which corresponds to the proton ^1H shifts (δ_H). The diagonal relationship of the signals in such 2D-NMR spectra is a direct consequence of the parallel between carbon chemical shifts and attached hydrogen shifts. The closely analogous Homonuclear shift correlation (HOMCOR) or COSY (COrrelated Spectroscop Y) maps out all homonuclear couplings in a single spectrum. In a typical COSY spectrum of stevioside (Fig. 15), both the F_1 and F_2 axes correspond to ^1H chemical shift scales. The normal ^1H NMR spectrum appears along the diagonal, while all off-diagonal signals corresponds to ^1H-^1H homonuclear couplings.

The HOMCOR technique was originally used by Jenner in 1971, when he obtained a COSY spectrum of β-chloroacrylic acid. It was only a decade ago that 2D-NMR became a routine technique.

a. Applications of 2D-NMR spectroscopy. Adequate COSY spectra were obtained from only 90 μg of cynarin (isolated from artichokes), allowing the distinction between two possible structures to be made. The complex structures of griseochelin and grisorixin metabolites, which are antibiotics obtained from strains of Streptomyces, were solved by COSY techniques. HETCOR experiments appear for the most part to have been used as a simple tool for assignment in the area of natural products. In the studies of bile acids, isomugineic acid (a metabolite derived from barley), the heteronuclear correlation was used to transfer proton assignments that had been made by other means onto the ^{13}C spectrum. HETCOR techniques have also been applied in long-range experiments to obtain structural information; for example, gracilin B (a bisnor-terpene isolated from the sponge Spongionella gracilis), ekeberginine (an alkaloid extracted from Ekebergia senegalensis), and didehydroryanodine (an insecticide obtained from Ryania speciosia) were all determined with the aid of long-range proton-carbon correlations using the HETCOR method.

Perhaps one of the recent exciting and fruitful areas of FTNMR research is in biochemistry. It is only been

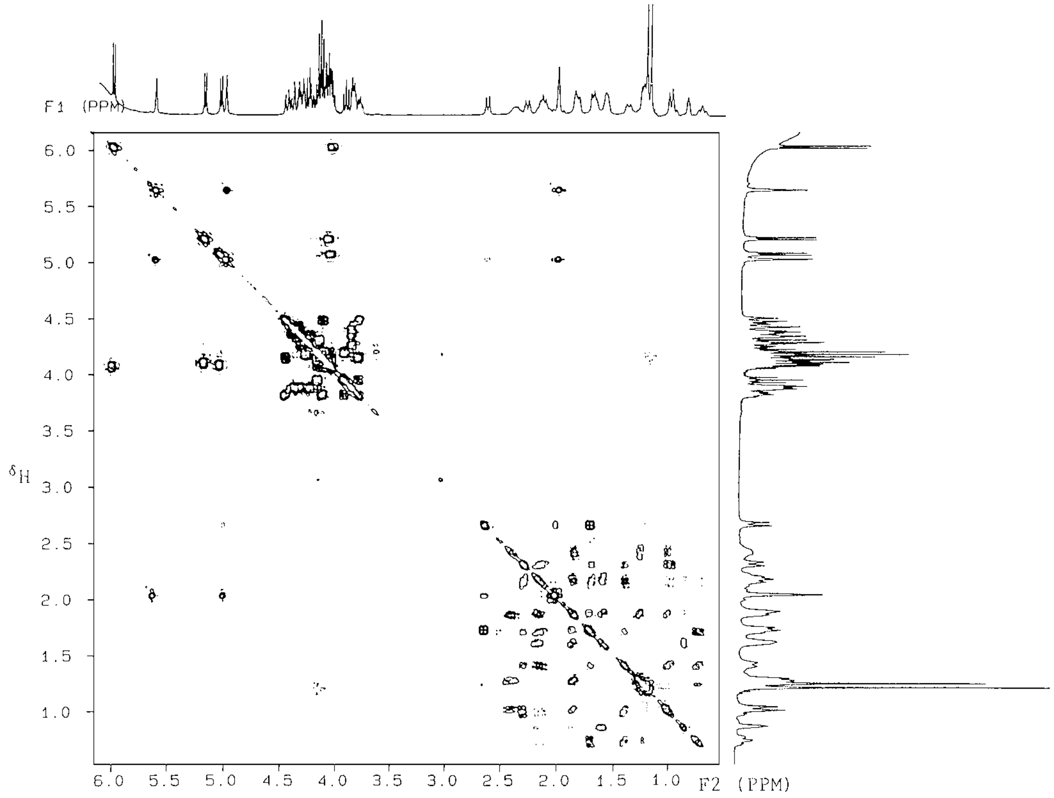


FIGURE 16 COSY NMR spectrum of stevioside (contour plot).

a decade since Moon and Richards at California Institute of Technology and Hoult at Oxford demonstrated that FT–^{31}P NMR spectroscopy is an effective probe of the *in vivo* generation and utilization of phosphate-bond energy. FTNMR spectroscopy has since been applied to the study of many diverse systems, progressing from intact cells to perfused organs to animals and man. *In vivo* FTNMR is noninvasive, and thus event-induced changes in the concentration of metabolites can be followed over a time period. ^{31}P is by far the most widely used nucleus for biochemical FTNMR determinations, as the key metabolite, ATP, adenosine diphosphate (ADP), creatine phosphate, and sugar phosphates, for example, all contain phosphorus. The NMR sensitivity of phosphorus is high (100%) and offers selectivity not obtained with nuclei such as ^{13}C or ^{1}H because it is not ubiquitous, but rather occurs primarily in cellular metabolites. Metabolic pathways are now being followed by ^{31}P FTNMR. Techniques using whole animals (or humans) require that the NMR measurements be localized to a particular organ or to a certain portion of the body. More recently is the application of FTNMR in the study of the metabolism of ^{13}C-labeled drugs. The fate of the labeled carbons can be followed as a function of time using ^{13}C NMR spectroscopy. Also new pulse sequences have now been developed to observe only those protons that are attached to the labeled carbons.

J. ^{1}H NMR "Polarimetry"

Optical purity of enantiomers is of great importance in the chemistry of natural products, especially in the fields of flavors, fragrances, and pheromones. Since odor receptors in mammals and in insects can discriminate between enantiomers, it is important to know the enantiomeric composition of chiral molecules that impinge on these receptors.

Determination of the enantiomeric composition of natural semiochemicals by ^{1}H NMR spectroscopy using a chiral lanthanide shift reagent is a rapid, accurate, and iondestructive technique. It is also a quantitative method for detection of adulteration of essential oils and oleoresins by synthetic racemic chemicals.

K. Supersonic Jet Spectroscopy

Capillary column gas chromatography (GC) is the highest resolution separation technique, but it is ultimately limited by the volatility of the molecules in the sample.

HPLC extends the range of species which can be carried in the analyte stream, but it suffers from relatively low column efficiency. Chromatography by itself provides extremely limited qualitative information. The most productive approach is the combination of an instrument having high separating power and one having high discriminating power. One measurement method with high discriminating power is supersonic jet (SJ) spectroscopy. Kantrowitz and Grey originally proposed supersonic expansions as a means of increasing the intensity of molecular beams in the early fifties, but it was in the mid-seventies that the SJ techniques were popularized for spectroscopic studies. It was only in the eighties that the analytical application to the study of complex samples were recorded. The combination of SJ spectroscopy with chromatography promises to be a powerful method for detailed analysis. Among all the line-narrowing spectroscopic techniques, SJ spectroscopy is the one which can be most directly coupled to chromatography. Advantages of the SJ approach include elimination of all matrix interferences, such as spectral shifts, site inhomogeneous broadening, phono-coupling, all other extrinsic fluorescence quenching, and Raman scatter and background from the matrx. The need to identify a suitable matrix is avoided. No complex cryogenic equipment is required. A disadvantage of SJ is the very short time which sample molecules spend in the observation volume, limiting the amount of signal averaging which can be done if only a small quantity of a sample is available or if it is to be detected in the narrow time span during which it elutes as a chromatographic band. Also, although the spectroscopic detection method may be nondestructive itself, the sample is usually lost. Among all the line-narrowing spectroscopic techniques, supersonic jet spectroscopy is the one which can be most easily used for real-time detection in chromatography.

The equipment required for SJ detection is less expensive than that for either MS or MI. Furthermore, the SJ approach allows for adding on downstream detectors, such as especially MS; the MS can be used in 100% and even weakly absorbing molecules should be detected.

1. Gas Chromatography–Supersonic Jet Spectroscopy (GC–SJ Spectroscopy)

It was in 1982 that the analytical potential of SJ spectroscopy was used to selectively detect the components of a mixture of naphthalene and α- and β-methyl naphthalenes. Trace impurities of fluorene in biphenyl at less than 1 ppm could be detected. Once the analytical advantages of SJ were realized, it would seem that coupling it to GC would be a narrow extension. A 100-ng detection has been recently achieved for perylene, and it has been estimated that 1-pg detection limit can be attained.

XIII. X-RAY CRYSTALLOGRAPHY

The purpose of X-ray analysis of crystal structures is to provide information on the positions of the individual atoms of a molecule, their interatomic distances, bond angles, and other features of molecular geometry such as the planarity of a particular group of atoms, the angles between the planes, and torsion angles around the bonds. The resulting three-dimensional representation of the atomic contents, rapidly determined by modern computerized techniques, of the crystal establishes the complete conformational structure and geometrical details hitherto unknown.

This information is of primary interest to chemists and biochemists who are concerned with the relation of structural features to chemical properties. It was only X-ray analysis that finally determined the double helix of two DNA molecules held together by hydrogen bonds that verified Watson and Crick's proposed model for DNA. Thus, X-ray analysis of organic compounds provides an unambiguous complete three-dimensional picture of the molecule, whereas other chemical and physical methods of structure analysis involved different aspects which collectively one can deduce the number and nature of the atoms bonded to each atom present (the topology of the molecule) or, for relatively simple molecules, provide some quantitative information from which geometrical details can be derived. There are instances whereby only X-ray analysis will decide absolutely the spatial arrangement of groups of atoms in simple organic molecules. It was X-ray analysis that confirmed that the cyclic sulphite (Fig. 17) had a boat–chair conformation with the S=O group trans to the cyclopropane ring and in an equatorial position.

The apparent disadvantages of X-ray analysis is that it is time consuming even with the use of automated computerized techniques. Further-more, X-ray analysis can only be carried out on pure crystalline materials. The crystals must be a certain size and regularity in order that a good X-ray structure of the compound is obtained.

XIV. CHEMICAL ANALYSIS

The requirement for information on chemical composition extends into most sectors of industry. The chemical measurement analytical market is extremely diverse. Chemical analysis is a key process in drug discovery, in assessing the quality of feedstock materials and end products, in monitoring chemical processes, in demonstrating regulatory compliance, in disease diagnostics, and in environmental conservation.

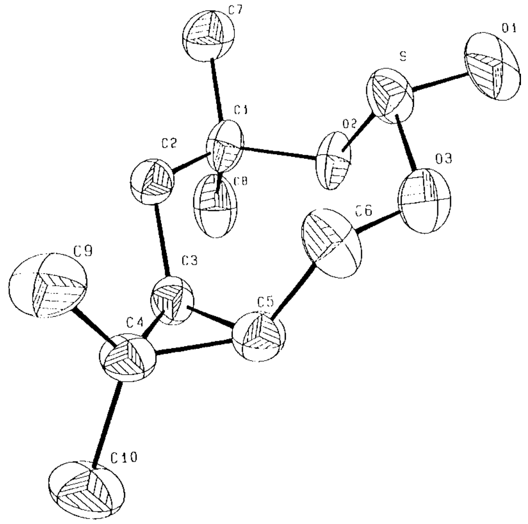


FIGURE 17 Molecular structure of eight-membered cyclic sulfite as revealed by X-ray analysis.

The majority of chemical measurements is currently carried out in dedicated analytical laboratories, where the samples are separated into their chemical components for subsequent detection and quantification. One of the main benefits of laboratory analysis is the reliability and accuracy that can be achieved.

Measurements performed using selective detection are reasonably reliable when chemical standards are available for quality control purposes. In recent years, however, stand-alone selective detectors in the form of chemical sensors have been considered as a means for performing measurement at-site.

To allow increased sample throughout and reduced analysis costs, the multitask approach to analysis is, in many areas, being superseded by methods that require fewer steps and allow greater parallel processing. Usually these methods include a selective detection stage, which then obviates the need for conventional purification and separation procedures. Selectivity in detection can be achieved by utilizing biomolecules (antibiotics, binding proteins, and enzymes) that exhibit high specificity for their target analytes. The use of immunoassays is now commonplace in clinical laboratories, enabling faster medical diagnosis for the patient.

There are now commercially available versions of the so-called "chemical nose" where multiple sensing elements are used to pick out specific chemicals in the aroma commodities such as beverages and perfumes. Most microanalytical techniques have been reproduced at a research level in an on-chip format. These include the selective detection systems, such as field effect transistors, ion-selective electrodes, biosensors, bioassays, and acoustic wave mass dectors.

XV. FUTURE PROSPECTS

The combination and computerization of spectroscopic and chromatographic methods is more efficient for the detection and identification of organic compounds than either of the two methods individually. The rapid development of NMR spectroscopic techniques, as well as the advent of even more powerful machines up to 600 MHz, is now being applied in the fields of medicine and biotechnology. The use of NMR spectroscopy in clinical diagnostics in the study of body fluids such as urine, sweat, amniotic fluid, and blood plasma is only just being realized. The application of NMR to biotechnology has tremendous potential, as 2D-NMR COSY spectra of proteins such as lysozyme have already been produced. It is envisaged that the last decade of the twentieth century will see *in vivo* monitoring and analysis of natural and genetically engineered biomolecules.

In his book *Philosophy of Sciences* published in 1836, Andre Marie Ampere made use of the strange term *cybernetique*, which in its modern concept denotes the multiscient robot of self-organizing machine. It is reasonable to assume that computer-controlled *cybermachines*, such as the coupled instruments, will soon become essential tools in every modern analytical laboratory.

It seems that the rapid development of the new techniques for the detection and identification of organic molecules (including biomolecules) is in accordance with the idea of the ancient Roman historian Tacitus, who said: *"Piscus crastinus papylo qui sapientam hodiernam continet,"* meaning "Tomorrow's fish is wrapped in the paper which contains today's wisdom."

SEE ALSO THE FOLLOWING ARTICLES

Gas Chromatography • Infrared Spectroscopy • Liquid Chromatography • Mass Spectrometry • Nuclear Magnetic Resonance (NMR) • Radioactivity • Raman Spectroscopy • X-Ray Analysis

BIBLIOGRAPHY

Afeyan, N. B., Fulton, S. P., Gordon, N. F., Mazsaroff, I., Váredy, L., and Regnier, F. E. (1990). "Perfusion chromatography—An approach to purifying biomolecules," *Bioltechnology* **8,** 203–206.

Agarwal, B. K. (1989). "X-Ray Spectroscopy," Springer, New York.

Alemany. L. B. (1989). "Exceptional resolution and new signals detected in the ^{13}C NMR spectra of alkanes," *Magnetic Resonance in Chemistry* **27,** 1065–1073.

Andrews, D. L., and Davis, A. M. C. (Eds.) (1995). "Frontier in Analytical Spectroscopy," The Royal Society of Chemistry, London.

Cooper, J. W. (1980). "Spectroscopic Techniques for Organic Chemists," Wiley, New York.

Craston and Cowen, S. (1998). "The Laboratory on a chip," *Science Progress* **81,** 225–244.

Duddeck, H., and Dietrich, W. (1989). "Structure Elucidation by Modern NMR," Springer, New York.

Firth, S. (1988). "Applications of diffuse reflectance infrared spectrometry," *Int. Lab.* **18**(8), 24–30.

Freemantle, M. (1999). "Downsizing chemistry," *Chem. Eng. News* Feb. **22,** 27–36.

Goates, R. G., and Sin, C. H. (1989). "Supersonic jet spectroscopy coupled to chromatography for very high resolution chemical analysis," *Appl. Spectrosc. Rev.* **25**(2), 81–126.

Granger, P., and Harris, R. K. (Eds.) (1991). "Multinuclear Magnetic Resonance in Liquids and Solids—Chemical Applications," Kluwer Academic, Dondrecht.

Griffith, P. R., and Haseth, J. A. (1986). "Fourier Transform Infrared Spectroscopy," Wiley, New York.

Grob, R. L., and Kaiser, M. A. (1982). "Environmental Problem Solving Using Gas and Liquid Chromatography," Journal of Chromatography Library, Vol. 21. Elsevier.

Gudzinowicz, B. J., and Gudzinowicz, M. J. (1979). "Analysis of Drugs and Metabolites by Gas-Chromatography–Mass Spectrometry," Chromatographic Science Series, Vol. 6, Marcel Dekker, New York.

Hayes, J. M. (1987). "Analytical spectroscopy in supersonic expansions," *Chemical Rev.* **87,** 745–760.

Heftman, E. (1983). "Chromatography," Journal of Chromatography Library, Vol. 22B, Elsevier.

Ikan, R. (1982). "Chromatography in Organic Microanalysis," Academic Press, New York.

Irwin, W. J. (1982). Analytical Pyrolysis, "Chromatographic Science Series," Vol. 22. Marcel Dekker. New York.

Levy, G. C., Lichter, R. L., and Nelsen, G. L. (1980). "Carbon-13 Nuclear Magnetic Resonance Spectroscopy," Wiley, New York.

McLafferty, F. W. (1983). "Tandem Mass Spectrometry," Wiley, New York.

Meuzelaar, H. L. C. (Ed.) (1990). "Computer Enhanced Analytical Spectroscopy," Plenum, New York.

Pryde, A., and Gilbert, M. T. (1979). "Applications of High Performance Liquid Chromatography," Chapman and Hall, London.

Schriemer, D. C., and Li, L. (1996). "Detection of high molecular weight narrow polydisperse polymers up to 1.5 million daltons by MALDI mass spectrometry," *Anal. Chem.* **68,** 2721–2725.

Silverstein, R. M., Bassler, G. C., and Morrill, T. C. (1991). "Spectrometric Identification of Organic Compounds," Wiley, New York.

Wilson, M. A. (1987). "NMR Techniques and Applications in Geochemistry and Soil Chemistry," Pergamon, Elmsford, New York.

Wilson, N. K., and Childers, J. W. (1989). "Recent advances in the matrix isolation infrared spectrometry of organic compounds," *Appl. Spectrosc. Rev.* **25**(1), 1–61.

Photoacoustic Spectroscopy

Konka Veeranjaneyulu
Roger M. Leblanc
University of Miami

I. Introduction
II. Photoacoustic Effect and Signal Generation
III. Advantages
IV. Instrumentation
V. Applications

GLOSSARY

Background light Nonmodulated white light used to saturate the photochemistry of the photosystems.

Depth-profile Depth specific characterization of a sample based on the frequency dependence of the thermal sample.

Modulation frequency The amplitude of the measuring light beam chopped at required frequency to generate the photoacoustic signal at the same frequency.

Optical absorption length Penetration depth of the incident light flux.

Phase angle Phase angle shift of photoacoustic signal relative to the modulation frequency.

Thermal deactivation Relaxation pathway of an optically excited molecule, wherein the excitation energy is converted into heat.

Thermal diffusion length The thermal diffusion length (μ_s) is defined by $\mu_s = \sqrt{\frac{\alpha}{\pi f}}$ where α is the thermal diffusivity of the sample and f is frequency.

Thermal diffusivity The thermal diffusivity is given by $\alpha = \frac{\kappa}{\rho c}$ where κ is the thermal conductivity, ρ is the density, and c is the thermal capacity.

This work is dedicated to Professor Danuta Frackowiak, Institute of Physics, Poznan University of Technology, Piotrowo 3, 60-965 Poznan, Poland, on the occasion of her 75th birthday.

THE PHOTOACOUSTIC effect is based on the conversion of absorbed light energy into heat by means of radiationless deexcitation processes. On irradiation of a solid or liquid sample by intensity modulated light, a heat wave of the same frequency is generated in the sample. The heat wave is transferred by diffusion to the surface of the sample, where oscillatory thermal effects are generated in the coupled gas. These can be detected as an acoustic signal. This technique is known as gas-coupled photoacoustic spectroscopy.

I. INTRODUCTION

Electromagnetic radiation, including the frequency range that is visible to the human eye, is a convenient and effective probe to study a wide range of biophysical and biochemical processes of biological systems.

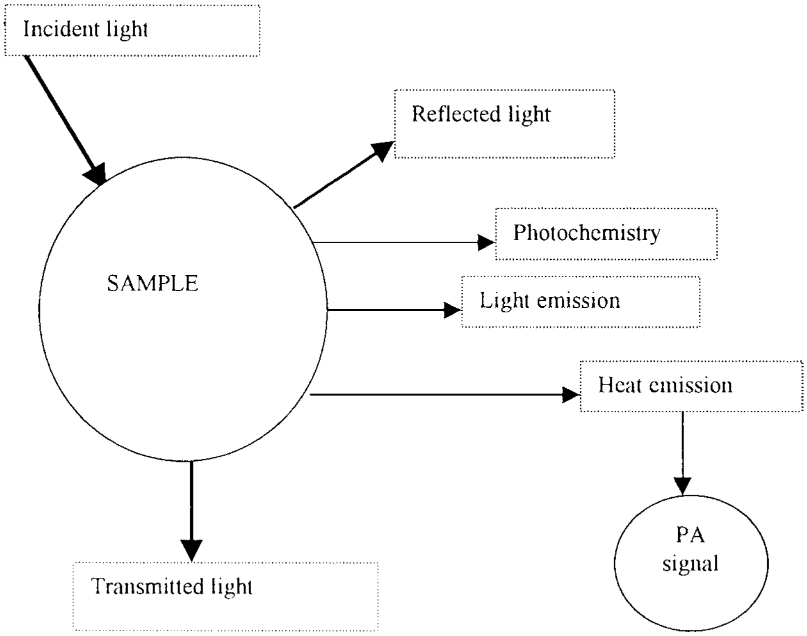


FIGURE 1 Various possible pathways of dissipation of incident electromagnetic radiation from a sample.

Ultraviolet(UV)/visible radiation is used to probe the electronic transitions, while infrared and microwave radiations are used to probe the molecular motions such as vibration and rotation. The UV and visible radiations are often used to probe the energetics of a system. The probing light incident upon the sample may be transmitted, absorbed, or reflected by the sample. In order to probe the cellular energetics of the system, the light must be absorbed by the sample. The conventional methods such as absorption spectroscopy or reflection spectroscopy measure the absorbed and reflected light, respectively. In absorption spectroscopy, it is the transmitted light that is measured, and the absorbance is calculated by considering the incident light (Io) and transmitted light (I). However, certain conditions such as nonhomogeneity and the light-scattering effect of the samples eliminate the use of absorption spectroscopy in a wide range of samples. Besides, the cellular biophysical activity can be better probed by studying directly the absorbed light energy rather than the transmitted light energy. Fluorescence spectroscopy has been used widely to probe complex biological systems. As most molecules of biological interest have appreciable fluorescence quantum yield, and the fluorescence can be measured from the surface of the sample, this technique has found wide usage and is considered the most sensitive tool. The fact that all molecules are not fluorescent, and even some fluorescent molecules exhibit low quantum yields, makes this technique limited to fluorescence high yielding systems.

Generally, when a sample absorbs light energy, a part of the absorbed energy is used for photochemistry (if the sample is photochemically active), and the rest is emitted as heat and light (fluorescence from singlet state and/or phosphorescence from triplet state). The excited molecules transfer the energy to the neighboring molecules or release the energy by radiative and/or nonradiative de-excitation processes. The nonradiative de-excitation process results in heat emission, and the radiative process emits light at longer wavelengths than the absorbed radiation (Fig. 1). As mentioned, light emission is measured by fluorescence spectroscopy. The heat emission can be measured by photoacoustic spectroscopy (PAS). Since every compound that absorbs light energy emits thermal energy, this technique is highly useful to probe directly the energetic and absorption characteristics of systems of interest. One particular aspect of this technique is the probing of absorbed light and not the transmitted light.

Photoacoustic spectroscopy is essentially a technique that uses the ability of periodically modulated or pulsed light to produce sound upon absorption. The photoacoustic (PA) effect is the result of the transformation of the absorbed modulated light energy into heat energy, which in turn generates sound. As a sample absorbs the modulated or pulsed light energy, it emits the heat periodically, which leads to the cycles of expansion and contraction causing the generation of sound waves. These sound waves are detected by a sensitive microphone. There are also other means to detect the modulated heat emission, based on the measurements of temperature changes. For example, in photochemical beam deflection spectroscopy, spatial and temporal refractive index changes caused by heat modulation are followed. These are probed by a weak laser beam, using the resulting deviations in the beam direction.

Another example is photothermal radiometry, where the modulated heat is monitored by the corresponding modulation in the amount of infrared radiation emitted from the sample, sensed by an infrared detector.

The PA effect was discovered in solid samples at the end of the nineteenth century. This was first reported over one hundred years ago by Alexander Graham Bell in his attempt to produce a light-activated telephone. In Bell's original observation, chopped sunlight was focused onto a selenium photocell, and Bell's main interest was to detect the electrical changes in the selenium cell. Incidentally he found that sound at this chopping frequency can heard by a nearby observer. The modern photoacoustic experiments show striking similarities to this effect, and this effect has proven to be an important and sensitive technique for studying the interaction between matter and light.

During the past two decades there has been revival of interest in photoacoustic phenomena, and intense efforts have been made to revive and develop suitable instrumentation. With the advent of modern electronics, a laboratory instrument suitable for investigating the spectral properties of optically thick samples was designed during 1970s. Soon a theoretical treatment for understanding the basic principles of the PA effect was given in many reviews. In principle, the incident light beam transforms the sample into an excited state. The thermal deactivation process from the sample induces the changes in the temperature of the thin layer of the coupling medium surrounding the sample. The coupling medium could be atmospheric air or any desired gas. These heat-induced changes in the coupling medium can be detected by various means. The signal origin and propagation in different kinds of samples such as solids, liquids, and gases are better explained in several review papers. In biological samples such as leaves and human skin, the signal generation is a complex phenomenon. The excited chromophore molecules of the cell deactivate thermally inside the cell, and the resulting heat wave propagates through the cytoplasm to the cell surface and contributes to the signal generation. The medium at the cell surface could be intercellular air in the case leaves, or an aqueous medium in the case of animal or human tissues. This technique has been applied to different fields of research. This review mainly focuses on the instrumentation and its potential applications.

II. PHOTOACOUSTIC EFFECT AND SIGNAL GENERATION

In general terms, the PA effect can be considered as a three-step sequence involving (1) intermittent absorption of light in a sample from a periodically modulated beam, (2) heat production by the absorbed energy, and finally (3) generation of acoustic pressure oscillations due to the thermal expansion.

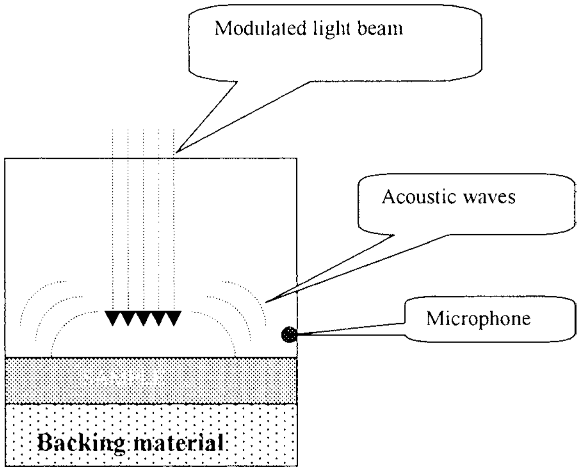


FIGURE 2 Photoacoustic signal generation depends on the thermal properties of the backing material, sample, and the gas inside the cell.

If a sample placed in a hermetically closed cell absorbs the modulated light of wavelength λ, the electrons will be excited to higher states. After a brief time, the electrons decay back to the ground state by radiative or nonradiative processes. The PA effect is solely due to nonradiative decay processes. A periodic heat distribution is created in the sample at the point of absorption. This thermal distribution then diffuses to the sample surface and gives up its thermal energy to the surrounding air (Fig. 2). A thin boundary layer of air responds to this heat, and it periodically expands and contracts as it is heated and cooled at the modulation frequency. This thin layer of air acts as an acoustic piston on the remaining column of air, creating pressure fluctuations which are detected by a microphone, and the resulting electronic signal is processed by a lock-in amplifier as the PA signal. Depending on the nature of the sample and experimental conditions, the acoustic signal generation may be within the sample or on the surface of the sample.

III. ADVANTAGES

The PA process has several important unique features which make PAS advantageous over conventional optical spectroscopy.

1. The PA signal is insensitive to the light-scattering effects of the sample. Unlike conventional optical spectroscopy, where scattered light gives rise to unwanted and misleading signals, in PAS only modulated absorbed light leads to a PA signal. Besides this, as the response is

modulated, only signals synchronized to the modulation are detected, and unwanted signals are eliminated.

2. Because it is insensitive to light scattering and light reflection properties of the sample (these features may reduce the light absorption and reduce the signal amplitude), this technique can be used to study the samples that cannot be studied by conventional spectroscopic methods.

3. This technique does not need any sample preparation; hence, the sample is not destroyed during the measurements. The sample can therefore be reused for repeated and time-bound investigations.

4. As the signal is the result of the thermal deactivation of the excited molecules of the sample, the PA spectrum represents the absorption spectrum of the sample.

5. As the thermal deactivation, fluorescence, and photochemistry (if present) are the competitive processes, the bioenergetics of the photochemically active intact samples can be investigated using this technique.

6. If the sample is photochemically active (such as green leaves), the product of modulated photochemistry (oxygen in the case of green leaves) also contributes to the signal. These compounds can also be characterized using the PA technique.

7. By suitably modifying the PA cell, the transmission spectrum of the sample can also be recorded using PAS.

8. PAS offers a unique feature of depth profiling of the compounds of interest in the sample. This important feature has led to its application in dermatological research, where drug penetration and solar cream diffusion in human skin have been investigated.

IV. INSTRUMENTATION

A. Modulated Photoacoustic Spectrometer

The essential components of a typical PA spectrometer are presented in Fig. 3. Light from a xenon arc lamp is passed through a monochromator for wavelength selection and is then modulated in intensity by a mechanical chopper whose frequency can be varied. The modulated beam is then focused on the sample placed in the PA cell. The acoustic waves are detected by a microphone, and the signals are sent to a lock-in amplifier and then to a computer or to a recorder.

B. Pulsed Photoacoustic Spectrometer

The pulsed PA setup consists of an Nd^{3+}:YAG laser, a beam splitter, a PA cell, an electronic interface, and a computer. The laser delivers the pulses of 3.5 nsec duration at 355 nm with typical energies of 1–2 mJ per pulse. The laser beam is split into two parts by a glass beam splitter, and a small percentage of energy (5–6%) is sent to a Joulemeter. This helps to correct the PA responses for any fluctuation of the output energy of the laser. The detection configuration of the PA cell is a classical one, in which the detector faces the irradiated surface of the sample. An integrated microphone located in front of the sample is used to detect the acoustic signal in the PA cell. The PA cell is a differential type, in the sense that it has two identical cylindrical cavities, one for the sample and another for the blank for the background noise measurements. Samples

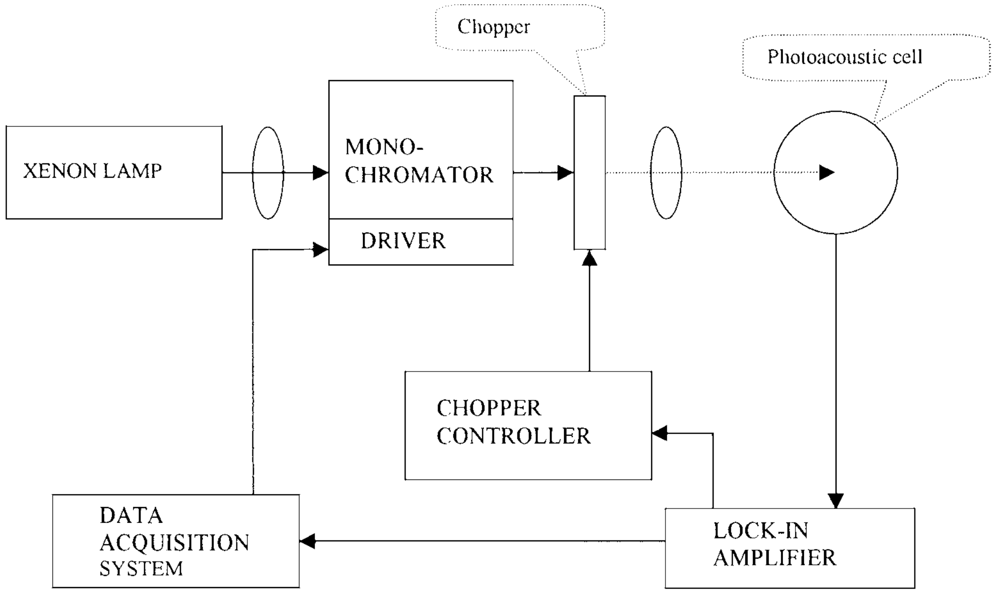


FIGURE 3 Schematic diagram of a typical modulated photoacoustic spectrometer.

are optically excited by groups of pulses, and data are acquired by a computer.

C. Photoacoustic Cells

During the past decade, PA cell architecture has transformed significantly from the basic concept of simple closed cell with a microphone to that of a highly evolved multichambered open cell capable of recording signals on live objects such as intact plants and human beings. Another factor that has improved significantly is the signal-to-noise ratio, where the cells were designed for taking measurements in the field. In the literature, one can see mention of a variety of photoacoustic cells.

1. Basic Photoacoustic Cell

The basic concept is that the PA cell must have a sensitive microphone, and it should be hermetically air-tight to generate an acoustic wave and to avoid external noise (Fig. 4). Usually the cell is fitted with a frontal quartz window for sending in the modulated excitation light onto the sample, and the volume of the cell should be small. Since it is air-tight, the gas composition of the cell can vary during the experiment. For example, a green leaf such as photosynthetically active green leaf, which takes up carbon dioxide from and releases oxygen into surrounding air, can alter the composition of the air in the cell. If one measures the photochemical activity of leaf, there may be limitation of the carbon dioxide to the photosynthetic phenomenon during the measurements. In order to avoid this

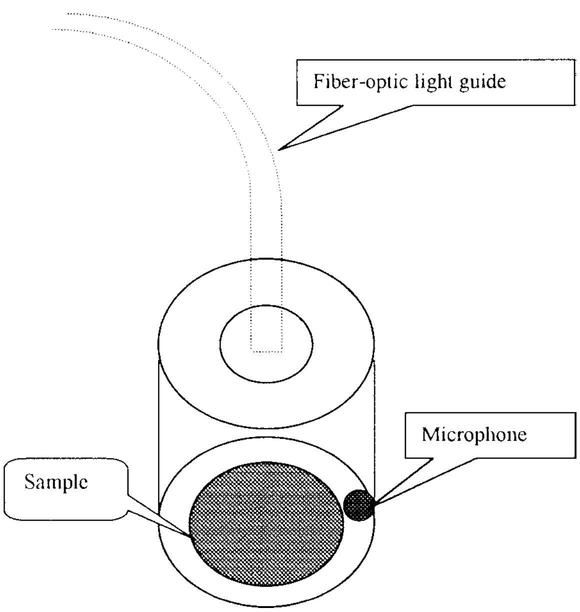


FIGURE 4 Components of a typical photoacoustic cell.

limitation, a gas-permeable cell was designed to enrich the carbon dioxide levels inside the closed cell. In addition, from the time the PA method was initially used, *in vivo* measurements were made on detached cut samples (either plant samples or animal or human skin samples). Though these measurements reflect the *in vivo* observations, realistically they may be called *in situ* observations. Hence, an open PA cell was designed that can be attached to the leaf, and is reported to measure the photosynthesis of a leaf still attached to the plant.

2. Photoacoustic Open Cell

Considering certain limitations with closed cells, investigators at different labs started working on developing open cells. Recently in our laboratory, we have designed a PA cell that can be attached directly to a live human being for measuring a drug or solar cream diffusion rates in the skin. The cell is designed to have identical cylindrical cavities fitted with microphones. A light pulse or beam is passed through a fiberoptic light guide. One microphone measures the background noise from the cardiac pulse, and the other detects the acoustic waves generated due to the excitation of compounds of interest inside the skin at the contact site. The cell can be strapped to any part of the body, and the signals can be recorded.

3. Cell with Optical Microphone

As mentioned above, the thermal deactivation-induced acoustic waves were detected by sensitive microphones in a majority of PA studies. In order to improve the sensitivity of the technique, a new detection system with optical microphone was used to detect the acoustic wave generated from the sample. In this system, the acoustic wave-induced change in the position of a laser beam on a pellicle is used to detect the signal. Instead of a microphone to detect the signal, a laser beam positioned on a Mylar pellicle (10 μM thickness) monitors the acoustic waves. The vibrations of the pellicle caused by acoustic waves deflect the laser beam, which is in turn is detected by a silicon photodiode (Fig. 5).

4. Cuvette Cell

A standard cuvette with 1-cm path length is used in this type of detection system. Samples are excited with a planar beam of light and the resulting pressure wave due to solvent expansion through the solution is converted by a transducer into a voltage pulse. As the measurements are made using a cuvette with an attached transducer, the system is easily adjustable to the needs of measurements in solutions.

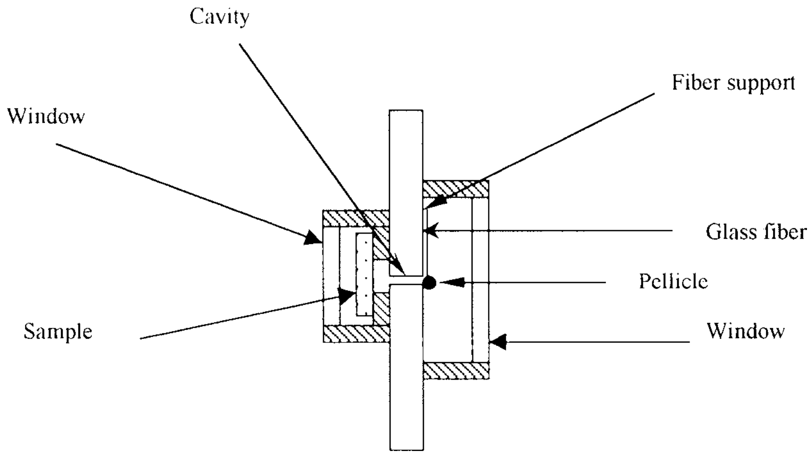


FIGURE 5 Schematic presentation of optical microphone.

5. Layered Prism Cell

Layered prism cell was assembled using a pair of dovetail prisms and a pair of piezoelectric transducers. The prisms are clamped together and separated by a shim. A 1-cm hole in the shim forms the sample compartment, and has the geometry of a thin disk. The inlet and outlet holes through one of the prisms permit the flow of the solution through the sample compartment. Light passes through the first prism, into the sample compartment, and out through the second prism. This cell uses the transmission of light through the prisms, and it combines the enhanced time resolution capabilities of "layered" front face irradiation geometry with zero-background and broadband flexibility of classical cuvette geometry. As described, different labs have developed different types of cells depending on the needs of their work.

V. APPLICATIONS

PAS proved to be a very useful and sensitive tool for photophysical and biophysical investigations in a number of fields. Since the early 1970s, several interesting observations have been made on organic molecules, gases, solids, isolated biomolecules, organelles, membranes, cell suspensions, microbial organisms, and intact leaves, among others. Various types of instruments have been used in collecting data on a variety of samples ranging from solids and suspensions to gels, liquids, and gases. Several *in vivo* studies have been made on plants, human objects, and microorganisms.

A. Chemical Analysis

PAS applications in chemical analysis have been widely reported, because of its unique features such as nondestructive, noncontact measurement, simple sample preparation, depth-profiling capability, and high signal saturation limit. It can be used for analyzing any kind of sample in any state. Thermophysical properties such as thermal diffusivity and conductivity of certain samples such as specialty papers have been investigated using Fourier transform-infrared–PAS (FTIR–PAS). This technique has been used to improve our understanding of the chemical origin of depth inhomogeneities.

FTIR–PAS can also be used as a rapid and nondestructive technique for quantitative characterization of the surface degradation of aged polymers and other compounds. Using FTIR–PAS spectroscopy, oxidation profiles of the various photooxidized polyisoprene–polystyrene copolymers have been investigated.

B. Biophysics of Proteins and Organic Molecules

The respiratory proteins hemoglobin and myoglobin have been extensively investigated by PAS for their structures in solution and in the crystalline state and for the mechanism of their interaction with binding substrates. In general it may be stated that PA calorimetry is a very important tool to investigate the changes in the extent of ligand formation, enthalpy, and volume characterizing the specific interactions of chromoproteins, including hydration and solvent interaction phenomena. Similarly, a series of proteins have been examined for their light-induced nonthermal volume changes by measuring heat release.

PAS has been of specific interest for investigating tryptophan because of its relevance as an intrinsic optical probe for proteins. Using PAS, intramolecular energy transfer processes have been investigated in crystal complexes of tryptophan with Sm(III), Tb(III), and Dy(III).

PA calorimetry has been successfully employed to understand the reactive intermediates of vitamin B_{12} compounds. It proved to be a direct method for the accurate

determination of bond dissociation energies in complex, biologically important molecules. It has also proved to be a sensitive indicator of conformational distortions and mechanistic detail as biological molecules undergo reactions. Using PAS, melanin content has been determined in human hair samples.

C. Dermatology

Skin forms a protective boundary layer between the body and its environment. With respect to transport properties, the nonvascularized epidermis, the outermost layer of which is formed by the *stratum corneum* (SC), serves as a barrier for the body. In the underlying vascularized dermis, the transport properties are strongly influenced by blood circulation.

As mentioned earlier, PAS is a good tool for studying the optical and thermal properties of a sample. Hence, this technique can be applied to fields of dermatological research such as drug detection and drug diffusion studies in skin, and thermal properties and water content of the skin. The optical and thermal properties of the sample are characterized by the optical absorption distance, μ_β, and the thermal diffusion distance, μ_s. The optical absorption distance is the inverse of the optical absorption coefficient β. The thermal diffusion length is defined by

$$\mu_s = \sqrt{\frac{\alpha}{\pi f}}$$

where α is the thermal diffusivity of the sample and f is frequency. The thermal diffusivity is given by

$$\alpha = \frac{\kappa}{\rho c}$$

where κ is the thermal conductivity, ρ is the density, and c is the thermal capacity. Note that the thermal diffusion length, μ_s, decreases with the increase in frequency and vice versa. Thus, by varying the modulation frequency, it is possible to study the diffusion rates of the drug in the skin.

Based on the ability to perform depth profile analysis, PAS has been shown to be suitable for the study of the penetration of substances topically applied to skin, in the visible spectral range. The absorption bands from specific group components of substances can be easily identified in the infrared spectral range, and a technique like PAS is highly advantageous for studying the distribution of substances through the skin. Using PAS, the depth profile of β-carotene and the differential penetration rate of β-carotene and methylene blue dye in the skin were investigated. It is reported that methylene blue dye stayed at the surface of the skin, while β-carotene penetrated up to the epithelium cell layer. Using FTIR–PAS, dimethylsulfoxide diffusion into the skin has been studied. Another example of drug detection in tissue using PAS is the detection of tetracycline topically applied to skin. Tetracycline was detected at 380 nm, where it strongly absorbs. The ease of identification of the substance in the spectral range suggests that penetration and distribution of other chemicals through the skin can also be investigated by FTIR–PAS.

A recent study reports a novel percutaneous absorptiometric system using an open-ended PA cell combined with a light guide. This system is capable of measuring the decrease of the amount of a drug over the skin *in vivo* at any given site with great ease, high sensitivity, and accuracy. Using the UV laser PA method, it demonstrates a series of *in vivo* and *in vitro* percutaneous absorptiometric measurements in the presence of anti-inflammatory and anti-itching agents. Melanin nonradiative de-excitation states and skin lipids have also been investigated in human skin samples.

Using PAS, the influence of SC constituents such as water, lipids, and hydrophilic substances on UV light absorption have been investigated. The data reveal that upon hydration of SC, there is a decrease of UV absorption coefficient by about 20%. Lipids were not shown to have any direct effect on UV absorption by SC. Further, the data suggest that the hydrophilic substances of SC were protected *in situ* by polar lipids enmeshed within the structure of SC.

Using classical modulated PA and pulsed PA techniques, skin hydration, drug penetration rates, and diffusion of solar creams into the human skin have been investigated. Using pulsed PAS (PPAS), diffusion kinetics of five solutions at different concentrations in a mixture of chromophores, as used in commercial sunscreens, were studied, and a two-step data interpretation for better characterization of diffusion characteristics was reported. The diffusion process is monitored by considering the amplitude of the heat response pulse and the time delay of maximum heat emission (Fig. 6). By analyzing the signal in

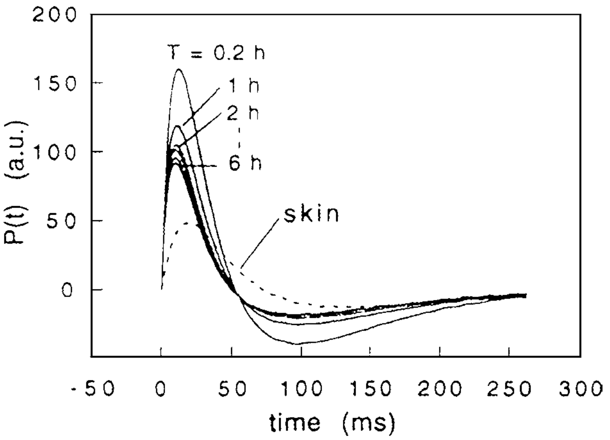


FIGURE 6 Pressure pulse evolution during diffusion kinetics for a 20% chromophore solution deposited onto the skin.

frequency and time domains, we found a clear difference in the diffusion behavior of the sunscreen compounds in the skin. The combination of Fourier transform combined with PPAS allowed us to determine the depth profile of the chromophores such as solar creams and other drugs. The frequency spectra helped to locate the chromophore diffusion with respect to the different layers of skin.

D. Oncology and Photosensitizers

Investigations on the photosensitized reactions in medicine, particularly in photodynamic therapy, have been of considerable interest in recent years. Intensive research efforts are being made currently to the study of photosensitized reactions in medicine, particularly in photodynamic therapy of tumors. Photodynamic therapy is a common treatment for cancer and various pathologies such as psoriasis. These treatments rely on the administration of photosensitizer and the irradiation of the zone treated. In the photosensitized reactions, the photosensitizer is excited by absorbed light and undergoes intersystem crossing to triplet state from which it either reacts with the nearby molecules or produces cytotoxic oxygen derivatives, such as singlet oxygen or superoxide anion. The drugs used for photosensitization processes must show very low fluorescent quantum yield and nonradiative decays with very high efficiency. In this sense, PAS is an excellent tool for studying the photophysical properties of the drugs and their interaction with biological molecules.

The photophysical properties of the photosensitizers are usually monitored to determine the efficiency of the treatments. Parameters such as the lifetime of excited states, the intersystem crossing yield, and singlet oxygen $O_2\,(^1\Delta_g)$ production yield are known only in solution form. As the PA and photothermal spectroscopic techniques are least sensitive to light scattering and are known to be sensitive to relaxation processes, several attempts have been made to use PAS to detect the photosensitizers in solution, in tissues, in excised organs, and in skin *in situ* and *in vivo*. Porphyrin derivatives are being actively investigated as photosensitizers for use in photodynamic therapy of tumors. The aggregation tendency of many porphyrins may be related to their selective retention in neoplastic tissues. Triplet state energy levels and intersystem crossing quantum yields of porphyrin derivatives have been obtained by measuring the phase angle difference between the PA signals of an unquenched sample and of the sample with triplet quenched. For molecules whose photophysical properties cannot easily be obtained by purely optical means, such an approach can prove to be a powerful method. Similarly it is also possible to study the reaction kinetics in environments that are not possible to monitor via absorbance. For example, when porphyrin based photochemotherapeutic agents are irradiated *in vivo*, a rapid photobleaching of the sensitizer is normally observed. PAS has been used to monitor such photobleaching in a highly scattering medium. It may also be possible to locate and delimit diseased areas in the case of external cancers from the hematoporphyrin signals and to follow the regression of the tumor. Similarly, photophysical parameters of the naphthalocyanines, merocyanine 540 having various applications in photobiology, have been investigated using fluorescence and PA spectroscopies. A significant improvement in instrumentation made during the past decade has resulted in much wider interest in the applications of PAS in dermatology.

E. Ophthalmology

The visual process is a series of reactions beginning with the absorption of a quantum of light energy. There are several temperature-dependent reactions during the conversion of the visual pigment rhodopsin to retinol. Several intermediates are formed during this process. Using the depth profile method, a three-dimensional image of the whole retina is recorded at low temperatures by varying the frequency. At high frequencies the rhodopsin–bathorhodopsin photostationary mixture appears to be the main absorbing species. At low frequencies, deeper layers are probed, and the PA spectra reveal the increasing amplitude of the PA band at 412 nm, which is due to blood hemoglobin. The PA data suggest that in the 25-μm layer which is probed between 50 and 60 Hz, the first half (0–13 μm) essentially contains the rhodopsin (outer rod segments).

The bovine retinal pigment epithelium (RPE) is a single layer of cells underlying the photoreceptor outer segments and is supported by the highly vascularized choroid. Upon dissection and removal of the retina, the RPE appears as an area of different colors, which are known as amelanotic and melanotic RPE. Using PAS, these amelanotic and melanotic RPE areas (Fig. 7A, B) are detected in the RPE. In healthy bovine eye lenses, PAS revealed the presence of the 340- to 360-nm band, which increases in magnitude with age and is particularly pronounced in brown nuclear cataracts. UV irradiation effects on eye lenses of dogfish sharks and rats have also been investigated using PAS.

Other biomedical applications of PAS include the study of the different states of blood hemoglobin, cytochromes, pharmaceutical powders, and diary products.

F. Plant Biology

1. Biomass Determination

PAS was used to determine the number of cells in the fermentation of cyanobacteria, as this technique is not

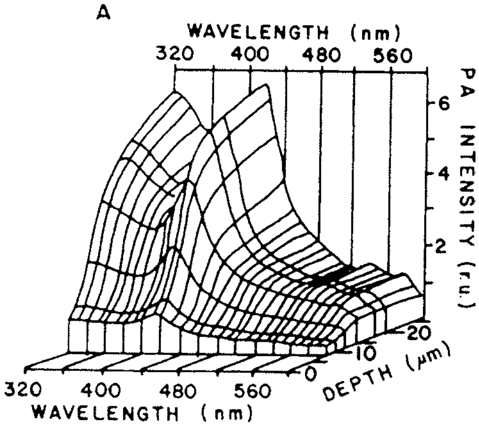


FIGURE 7A Photoacoustic spectra recorded at room temperature with bovine amelanotic pigment epithelium.

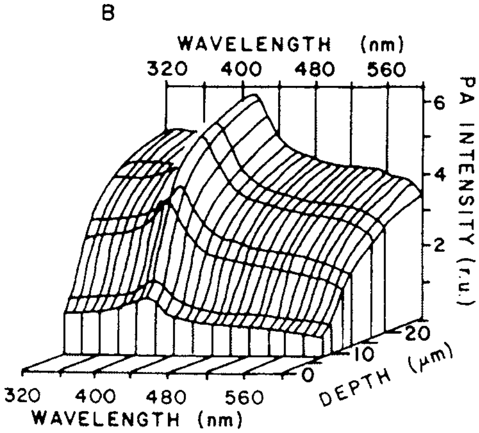


FIGURE 7B Photoacoustic spectra measured for the melanotic epithelium.

influenced by scattering particles and gas bubbles. PAS has proved to be very useful in the cases of very low and high number of cells; at these concentrations, the transmission was found to show pronounced deviation from the true number of cells.

2. Photosynthesis

Another important field where PAS has been successfully employed is green plant photosynthesis. Although there are many techniques available to study photosynthesis, not many are applicable to the comprehensive investigation of photosynthesis of intact plants. Even among the techniques that measure the intact leaf photosynthesis, their applications are limited to specific photosynthetic parameters. For example, an infrared gas analyzer can be used only to measure the carbon dioxide uptake during photosynthesis. Similarly, the oxygen electrode technique can help to measure photosynthetic oxygen evolution, and the fluorescence technique can be used to monitor the photochemistry of photosystem II. Unlike these techniques, PAS can be used to measure photosynthetic oxygen evolution, and to determine the photochemistry of photosystems I and II. During the past two decades, this technique has been applied to answer several questions of intact leaf photochemistry.

Several reviews have been published on the application of PAS to the photosynthesis of green plants. Using PAS, certain aspects of intact leaf photosynthesis, which cannot be understood by other techniques, have been studied. The study of photosynthetic molecular mechanism and structure is probably the most popular application of PAS.

a. Photoacoustic signal generation in green leaves. In all the above mentioned applications, we have been describing that the thermal deactivation of excited molecules from the sample alone contributes to the PA signal. However, in green leaves, besides the thermal deactivation of pigments, the product of the modulated photochemistry such as photosynthetic oxygen evolution also contributes to the signal depending on the modulation frequency. In plant cells, chloroplasts are the sites of photosynthesis. These organelles are suspended in the cytoplasm of the cell. Oxygen molecules generated during the water oxidation phenomenon of primary photochemical reactions have to travel to the surface of the cell in order to contribute to the acoustic signal. Thus, the photoacoustic signal from leaves is a composite of modulated heat and oxygen emissions depending on the modulation frequency. At low modulation frequency ($<$200 Hz), both heat and oxygen contribute to the signal, and at higher modulation frequencies ($>$200 Hz), the oxygen signal is dampened and only heat emission contributes to the signal. The oxygen signal can be dampened by infiltrating the leaves with water; and heat emission alone can also be measured even at low modulation frequencies.

The heat and oxygen emissions can be separated vectorially in the presence of a saturated strong background light using a lock-in amplifier in the two-phase mode by recording the signal in the in-phase and quadrature channels. The saturated background light closes the reaction centers, and all the absorbed modulated measuring light is released as heat without contributing to modulated photochemistry. In the presence of background light, the signal in one channel (quadrature) can be adjusted to zero by adjusting the phase angle. Upon switching off the background light, the modulated measuring light contributes to the photochemistry and the oxygen signal can be seen in the quadrature channel (Fig. 8). The amplitude of the oxygen signal (A_{ox}) can be evaluated according to the standard

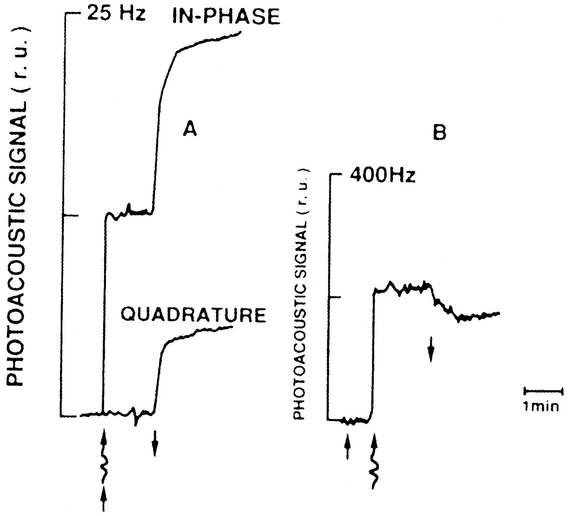


FIGURE 8 Photoacoustic signals from pea leaves. (A) Signal at 25 Hz in the absence of background light is vectorially separated into modulated oxygen evolution in the quadrature channel, and modulated heat and oxygen in the in-phase channel. (B) Signal at 400 Hz in the presence and absence of background light. Wavy arrow: modulated light; straight arrow: background light.

methods after correcting for the energy storage measured at higher (>200 Hz) frequencies by using the equation

$$A_{\text{ox}} = \sqrt{R_s^2 + (R_c - T_c' \times K)^2}$$

where R_s and R_c are the quadrature and in-phase signals, respectively, in the absence of background light, T_c' is the in-phase signal in the presence of background light (A_{PT}), and K is the ratio of acoustic signal in the absence of background light (Q_m) to the signal in the presence of background light (Q_{ma}) at high frequency. Thus, both heat emission and oxygen evolution from green leaves can be determined by recording the signal at different frequencies.

b. Photosynthetic energy storage. During the photochemical processes of photosynthesis, a fraction of absorbed light energy is stored as free energy in chemical intermediates. The magnitude of this stored energy depends on the quantum yield of the primary photochemical events and on energy levels of various intermediates and their decay rates. This photochemical energy can be measured by comparing the PA signal from a photosynthetically active sample with that of a photosynthetically inactive reference. This photosynthetically inactive reference can be obtained by saturating modulated photochemistry of the sample (self-reference) with a nonmodulated strong light beam. Thus by recording the signal in the presence (Q_{ma}) and in the absence (Q_m) of background white light, photosynthetic energy storage can be determined. This energy represents total energy stored by photosystem I (PS_I) and photosystem II (PS_{II}).

$$\text{ES}_\text{T} = \text{PS}_\text{I} + \text{PS}_\text{II} = \frac{Q_{ma} - Q_m}{Q_{ma}} \times 100$$

Both PS_I and PS_{II} absorb the most part of the spectral radiation in the visible range except in the far-red region, where PS_I alone absorbs. This was well demonstrated by measuring the PS_I associated activity using modulated far-red light (700 nm). Hence any energy storage measured using the modulated light in the spectral range between 400 and 695 nm in the presence of white background light reflects the energy stored by both PS_I and PS_{II}.

By using a nonmodulated strong background far-red light (>715 nm, which is absorbed only by PS_I), energy storage of PS_I can also be determined. The difference in the amplitude of the signal in the absence (Q_m) and in the presence (Q_{mfrl}) of background far-red light indicates the amount of the measuring light energy absorbed by PS_I. By subtracting the PS_I stored energy from total energy stored, energy stored by PS_{II} can be derived. This is possible only by using PAS.

$$\text{ES}_{\text{PS}_\text{I}} = \frac{Q_{mfrl} - Q_m}{Q_{ma}} \times 100$$

Then

$$\text{ES}_{\text{PS}_\text{II}} = \text{ES}_\text{T} - \text{ES}_{\text{PS}_\text{I}}$$

Using this model, clear and direct evidence was presented for light state transitions and migration of light harvesting complex II (LHC_{II}) between PS_I and PS_{II}.

c. Light state transitions. In green algae and higher plants containing chlorophyll b as major accessory pigment, PS_{II} absorbs more light at short wavelength region ($\lambda < 670$ nm) than does PS_I. PS_I alone absorbs in the far-red region of the spectrum ($\lambda > 715$ nm). Exposure of an intact leaf to shortwavelength light leads to adaptation of its photosynthetic apparatus to the state 2 condition by redistributing the energy in favor of PS_I to have a balanced excitation of both photosystems. This is reversible in the presence of far-red light, leading to the state 1 where short-wavelength light largely excites PS_{II}. The mechanism of this energy redistribution between PS_I and PS_{II} was under intense debate until the application of PAS to plant photosynthesis. Using PA oxygen evolution and complementary fluorescence measurements, it was demonstrated that LHC_{II} migrates between PS_I and PS_{II} in order to have balanced excitation distribution. Using the energy storage model as described above, it was convincingly demonstrated that LHC_{II} migrates between PS_I and PS_{II} and changes the absorption cross section of photosystems. The data presented in Fig. 9 and Table I formed

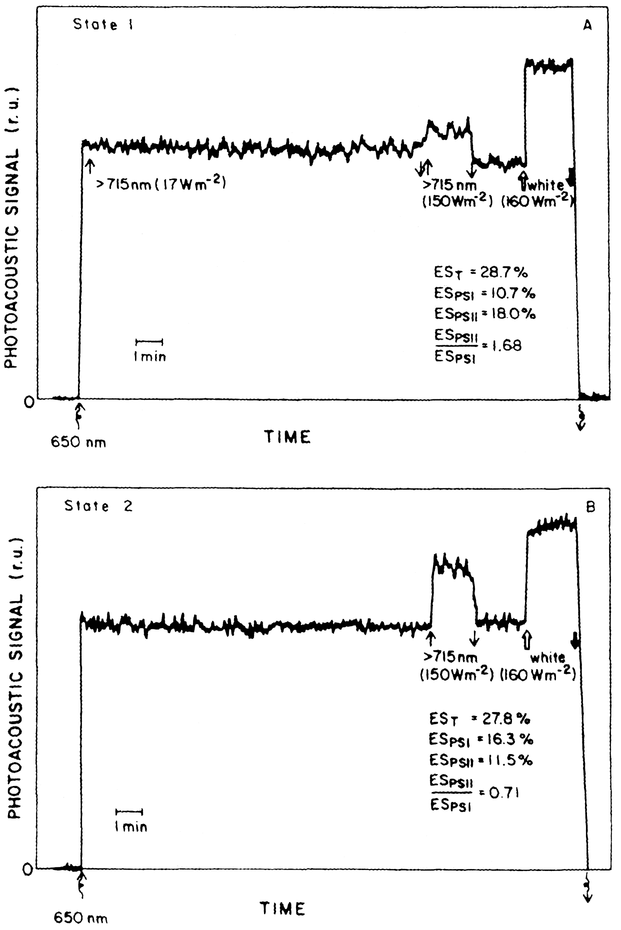


FIGURE 9 Photoacoustic signal from sugar maple leaf in the course of time during state 1 (A) and state 2 (B).

the direct evidence for the migration of LHC_{II} between PS_I and PS_{II}. Energy storage of both systems changes during state transitions. The fraction of energy gained by PS_I is nearly equal to that lost by PS_{II} during transition to state 2 and vice versa in state 1. This fraction amounts to about 5% of absorbed light, 19% of total energy stored, 34% of PS_{II} energy stored, and 43% of PS_I energy stored in state 2.

TABLE I ES_T, ES_{PS_I}, and $ES_{PS_{II}}$, Ratio of $ES_{PS_I}/ES_{PS_{II}}$, and Light Distribution between PS_I (α) and PS_{II} (β) during State 1–State 2 Transition in Sugar Maple Leaves

	State 1	State 2	(State 2 – State 1)
ES_T (%)	27.3 ± 3.7	26.9 ± 3.4	−0.4
ES_{PS_I}(%)	6.7 ± 4.0	11.8 ± 4.3	+5.1
$ES_{PS_{II}}$(%)	20.6 ± 5.5	15.1 ± 5.8	−5.5
$ES_{PS_I}/ES_{PS_{II}}$	3.1	1.3	−1.8
α	0.24	0.44	+0.19
β	0.75	0.56	−0.19

This was the first time such quantitative information on the amount of energy transfer between PS_I and PS_{II} was provided. All these measurements were made with 650 nm modulated light.

In order to understand these changes in the spectral range of photosynthetically active radiation, spectra were recorded under state 1 and state 2 conditions (Fig. 10). The photochemical activity remains the same in both state 1 and state 2 between 580 and 700 nm, but it is lower in state 2 between 400 and 580 nm. Both PS_I and PS_{II} activities change significantly due to the migration of LHC_{II}. The amplitude of LHC_{II} coupling-induced change is the same in both PS_I and PS_{II} between 580 and 700 nm, but it is less in PS_I than in PS_{II} between 400 and 580 nm, which explains the lower photochemical activity of the leaf in state 2 than in state 1. This is due to the decrease in energy transfer efficiency of the carotenoids to chlorophylls in LHC_{II} when it is associated with PS_I.

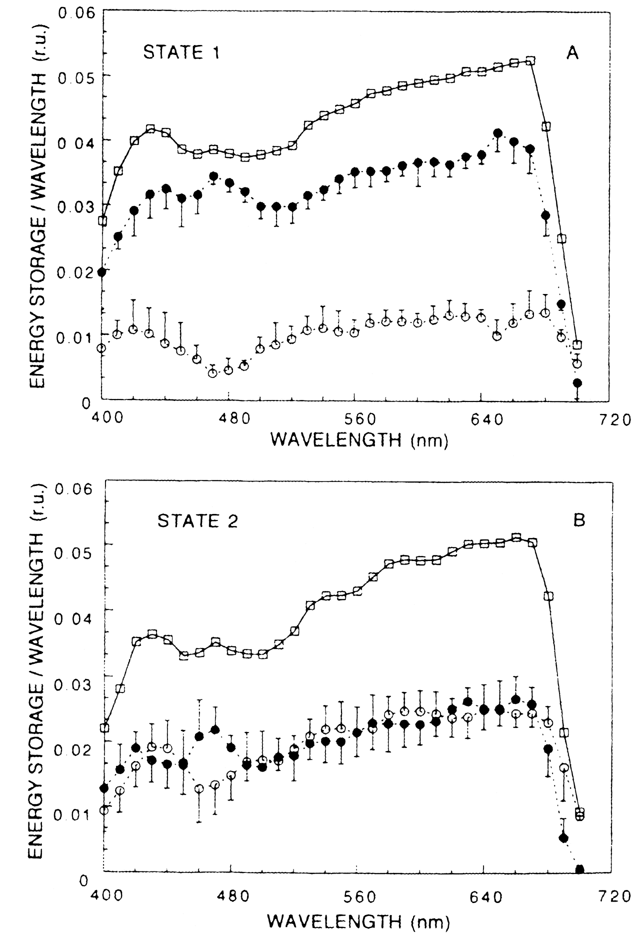


FIGURE 10 Energy storage of both photosystem I and II together (squares), photosystem I (open circles), and photosystem II (closed circles) of sugar maples leaves.

d. Photosynthetic oxygen evolution. In photochemical reactions of photosynthesis, water is the primary electron donor and NADP is the ultimate electron acceptor. During the process of water oxidation, oxygen is released as a by-product of photosynthesis. In the low-frequency range, PAS detects this oxygen produced during photosynthesis. By recording the signal and normalizing it with the heat signal, one can study the relative quantum yield of photosynthetic evolution.

By measuring energy storage, oxygen evolution, photosynthetic transients, light distribution between PS_I and PS_{II}, low light state adaptations, and independent activity of PS_I *in vivo* have been examined.

Several environmental stress effects on plants have been investigated using PAS. PA methods have been used to monitor the effects of water stress, high-light stress, low- and high-temperature stress, and chemical pollutant stress on plants. As the technique facilitates understanding the photochemical activity of intact leaves, several studies have been made to investigate drought stress- and high light stress-induced changes in the photochemical activity of the leaves. PAS also helps one to understand the light distribution and balanced excitation of photosytems during stress conditions. Under high light conditions, heat emission has been considered a protective mechanism. PAS was used to measure the heat emission in high light-treated plants, and it was demonstrated that plants dissipate the excess light energy in the form of heat. In shade-adapted sugar maple plants, an instantaneous heat emission upon exposure to high light was demonstrated.

Using PAS, the mechanism of action of bisulfite in pea leaves was investigated. Using oxygen free-radical scavengers, it was demonstrated that the Calvin cycle is a valid candidate for the primary site of action of bisulfite in plants. Using PA and fluorescence techniques, it was also demonstrated that abscisic acid offers protection against the inhibition of photosynthesis by bisulfite. Interestingly, the stimulatory effect of sulfur dioxide on maple leaf photochemical activity was also reported using PAS. Thus PAS has proved to be a versatile technique for studying environmental stress effects on plants.

3. Spectral Properties

As mentioned above, there is oxygen contribution to the PA signal in the low-frequency range. The amplitude of the photochemistry depends on the wavelength of excitation light. In order to avoid the spectral deformities due to modulated photochemistry, PA spectra of intact leaves are recorded in the presence of nonmodulated, strong, background white light, which closes the reaction centers and allows all the absorbed modulated light to escape as modulated heat. PA spectra of green leaves recorded under such conditions resemble those of absorption spectra (Fig. 11). PA spectra have been used to identify and characterize pigments and to analyze the depth profiles of pigments in intact leaves. The spectra have also been used to understand the physiological intactness of plants under environmental stress conditions, to understand the energy transfer process among pigment molecules, and to determine the relative quantum yield of photochemistry in spectral range of photosynthetically active radiation.

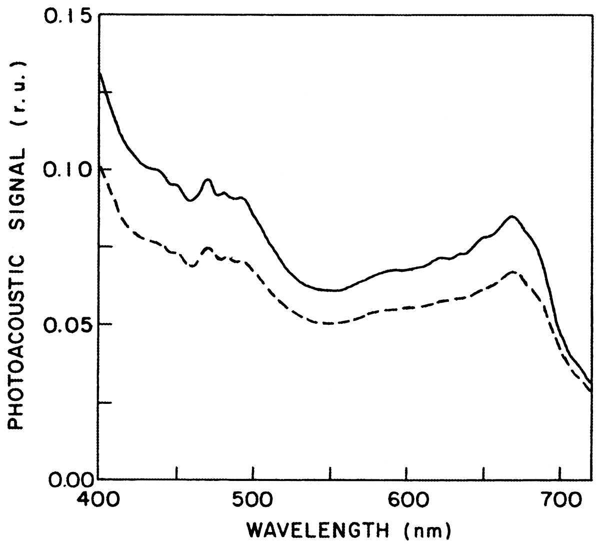


FIGURE 11 Photoacoustic spectrum of pea leaf recorded at 40 Hz in the absence (solid line) and presence (broken line) of background light.

G. Environmental Sciences

Laser PAS has been used to monitor air samples. Minimum detectable concentrations of gaseous pollutants are often in the parts per billion (ppb) or sub-ppb range depending on the molecular absorption cross section and on possible absorption interferences. Most studies have been devoted to investigations on collected air samples of different origin. A CO laser-based PAS has been used for analysis of vehicular exhausts. A number of compounds such as nitric oxide, nitrogen dioxide, H_2O vapor, and other volatile organic compounds such as alkenes, aromatic hydrocarbons, and aldehydes have been reported. A mobile PAS system has been developed for analysis in the field.

Similarly, analysis and time-resolved monitoring of stack emissions from power plants, incineration plants, or industrial plants are also of considerable interest. Air samples have been analyzed using the mobile PAS system in urban and rural areas. Thus, PA monitoring has not been restricted to laboratory measurements, but is performed in the field.

SEE ALSO THE FOLLOWING ARTICLES

ACOUSTIC WAVE DEVICES • AUGER ELECTRON SPECTROSCOPY • INFRARED SPECTROSCOPY • MODULATION • PHOTOCHEMISTRY, MOLECULAR • POLLUTION, AIR

BIBLIOGRAPHY

Autrey, T., *et al*. (1998). "A new angle into time-resolved photoacoustic spectroscopy: A layered prism cell increases experimental flexibility," *Rev. Sci. Instrum.* **69,** 2246–2258.

Bock, G., and Harnett, S. (1989). "Photosensitizing Compounds: Their Chemistry and Clinical Use," Wiley, Chichester, UK.

Boucher, F., and Leblanc, R. M. (1981). "Photoacoustic spectroscopy of cattle visual pigment at low temperature," *Biochem. Biophys. Res. Commun.* **100,** 385–390.

Braslavsky, S. E., and Heibel, G. E. (1992). "Time resolved photothermal and photoacoustic methods applied to photo-induced processes in solutions," *Chem. Rev.* **92,** 1381–1410.

Buschmann, C. (1990). "Photoacoustic Spectroscopy—Photoacoustic and Photothermal Effects," *In* "Modern Methods of Plant Analysis, New Series," Vol. 11, Springer-Verlag, Berlin.

Cahen, D., *et al.* (1980). "Photoacoustics in life sciences,"*J. Biochem. Biophys. Method.* **3,** 293–310.

Canaani, O., Barber, J., and Malkin, S. (1984). "Evidence that phosphorylation and dephosphorylation regulate the distribution of excitation energy between the two photosystems in vivo: Photoacoustic and fluorimetric study of an intact leaf,"*Proc. Natl. Acad. Sci. USA* **81,** 1614–1618.

Charland, M., and Leblanc, R. M. (1993). "Photoacoustic spectroscopy applied to biological systems," Bulletin of the Institute for Chemical Research, Kyoto University, **71,** 226–244.

Crippa, P. R., Vecli, A., and Viappiani, C. (1994). "Time-resolved photoacoustic spectroscopy: New developments of an old idea," *J. Photochem. Photobiol. B: Biol.* **24,** 3–15.

Fork, D. C., and Herbert, S. K. (1993). "The application of photoacoustic techniques to studies of photosynthesis," *Photochem. Photobiol*. **57,** 207–220.

De Paula, M. H., Vinha, C. A., and Badini, R. G. (1992). "High-sensitivity optical microphone for photoacoustics," *Rev. Sci. Instrum.* **63,** 3487–3491.

Malkin, S., and Canaani, O. (1994). "The use of the characteristics of the photoacoustic method in the study of photosynthesis," *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **45,** 493–526.

Repond, P., and Sigrist, M. W. (1996). "Photoacoustic spectroscopy on trace gases with continuously tunable CO_2 laser," *Appl. Optics* **35,** 4065–4085.

Rosencwaig, A. (1975). "Photoacoustic spectroscopy. A new tool for investigation of solids," *Anal. Chem.* **47,** 592–604.

Rosencwaig, A., and Gersho, A. (1976). "Theory of photoacoustic effect with solids," *J. Appl. Phys.* **4,** 64–69.

Sigrist, M. (1998). "Encyclopedia of Environmental Analysis and Remediation," Wiley, New York.

Veeranjaneyulu, K., *et al.* (1991). "Photoacoustic study of changes in energy storage of photosystems I and II during state 1–state 2 transitions," *Plant Physiol.* **34,** 271–334.

Photoelectron Spectroscopy

G. Hohlneicher
A. Gildenpfennig
University of Cologne

I. Fundamentals of Photoelectron Spectroscopy
II. Photoemission from Atoms, Molecules, and Solids
III. Instrumentation
IV. Some Examples
V. New Directions

GLOSSARY

Auger electron spectroscopy Investigation of the kinetic energy distribution of electrons produced through the decay of a highly excited core hole state.

Binding energy The energy necessary to remove an electron from a certain orbital or shell.

Characteristic X-ray emission Emission of X-rays with well-defined energy resulting from the radiative decay of core hole states.

Chemical shift Variation in the binding energy of an inner shell electron due to different chemical environments.

Core hole state Excited state of an atom in which an electron has been removed from one of the inner shells.

Multiplet splitting Substructure of bands in photoelectron and Auger electron spectra caused by different spin states and by spin–orbit interaction.

Photoelectron spectroscopy Investigation of the kinetic energy distribution of photoemitted electrons.

Photoemission Emission of electrons from a free molecule or a surface following irradiation with photons of sufficiently high energy.

Vibrational fine structure Structure observed in highly resolved photoelectron spectra due to interaction of the electronic process with molecular vibrations.

PHOTOELECTRON SPECTROSCOPY is the investigation of the energy distribution of electrons released from a sample by irradiation with sufficiently energetic radiation. It is applied to a wide range of materials, such as atoms or molecules in the gas phase, solids, and—with very special technical requirements—liquids. Photoelectron spectroscopy is a powerful analytical tool for the investigation of the electronic structure of molecules and solid surfaces. It has become especially important in surface science because of its extraordinary surface sensitivity.

I. FUNDAMENTALS OF PHOTOELECTRON SPECTROSCOPY

A. Basic Principles

As long as there is no need to specify the nature of the sample, we simply speak of the investigated system (M).

Before photoionization takes place, the system is in a well-defined electronic state, usually the electronic ground state M_0, the initial state of the photoemission process. Irradiation of the system with radiation of sufficiently high energy $h\nu$ leads to the ejection of a photoelectron. The ion M^+ that is created by this process is again in a well-defined electronic state M_i^+ which is the final state of the photoemission process. In general, the lifetime of M_i^+ is long enough to prevent successive changes in the ion state from influencing the kinetic energy of the photoelectron. The final state is either the electronic ground state M_0^+ or, provided $h\nu$ is sufficiently high, an electronically excited state of the ion. From conservation of energy it follows that

$$E(M_i^+) - E(M_0) + E_{kin}(e^-) + E_{kin}(M^+) = h\nu \qquad (1)$$

where $E(M_0)$ and $E(M_i^+)$ are the energies of the initial and final states, respectively, and $E_{kin}(e^-)$ and $E_{kin}(M^+)$ are the kinetic energies of the electron and of the ion. Since $E_{kin}(e^-)/E_{kin}(M^+)$ is determined by the mass ratio $m(M^+)/m(e^-)$, $E_{kin}(M^+)$ can be neglected in most applications. This leads to

$$E_B(i) = E(M_i^+) - E(M_0) = h\nu - E_{kin}(e^-) \qquad (2)$$

where $E_B(i)$ is called the "binding energy." To avoid confusion, this expression is used throughout this article. However, specifically in connection with the investigation of free molecules, the energy difference $E(M_i^+) - E(M_0)$ is also referred to as the "ionization energy" or "ionization potential."

If the excitation energy $h\nu$ is known, Eq. (2) allows determination of $E_B(i)$ from the kinetic energy of the created photoelectrons. Photoelectron spectroscopy (PES) is basically the measurement of the kinetic energy of photoelectrons with the goal of deriving information about binding energies. This definition distinguishes PES from other methods in which photoionization is used mainly for detection (e.g., laser-induced multiphoton ionization) and not for determining binding energies.

B. Photoelectron Spectra

A photoelectron (PE) spectrum is the number of photoelectrons with kinetic energy $E_{kin}(e^-)$ observed per unit time, displayed as a function of kinetic energy. Three examples are shown in Fig. 1: The gas-phase spectrum of neon excited with $h\nu = 1253.6$ eV (Fig. 1a), the gas-phase spectrum of $H_2C{=}O$ excited with $h\nu = 21.2$ eV (Fig. 1b), and the solid-state spectrum of copper excited with $h\nu = 1486.7$ eV (Fig. 1c). The meaning of the assignments given in these spectra is explained below. In all three cases two energy scales are shown corresponding to the IUPAC recommendations: The scale for the kinetic energy is given at the top of the spectrum and the scale for the binding energy at the bottom. The binding energy scale is obtained from the kinetic energy scale by means of Eq. (2). It must be kept in mind, however, that the quantity originally measured is $E_{kin}(e^-)$, even when only a scale for E_B is shown. The scales in Fig. 1 run in different directions for different examples. This is not a mistake but is due to different presentations of PE spectra in the literature. Some researchers show values of the measured quantity $E_{kin}(e^-)$ increasing from left to right, which results in the scale for E_B running from right to left. This kind of display is especially common for PE spectra of solid surfaces. Others are interested mainly in E_B, so they draw the E_B scale with increasing energies from left to right. Sometimes only the $E_{kin}(e^-)$ scale or only the E_B scale is provided. As far as possible we will show both scales throughout this article, but when using spectra from the literature one should always be aware of the scale used.

Figure 1 gives a first impression of different types of PE spectra. From Eq. (2) we expect photoelectrons to appear only at kinetic energies that correspond to a certain final state M_i^+. We therefore expect a PE spectrum to consist of lines with widths defined by some experimental parameters. This is the result observed for neon (Fig. 1a). For molecules, vibrational and rotational states are coupled to the electronic states and, as in optical spectroscopy, band spectra are obtained (Fig. 1b). In this case we speak of photoemission bands rather than photoemission lines. With the exception of the technique discussed in Section I.I, the resolution achievable in PES is much lower than in optical spectroscopy (Section III.B). If the sample is a solid (Fig. 1c), each band or line is preceded by a tail extending toward lower kinetic energies. For low kinetic energies this leads to a considerable background. The tails are due to inelastic scattering of the photoelectrons within the solid.

The progress achieved with PES results from the fact that it is an energy-resolved method. Older methods for the determination of binding energies were mainly based on a measurement of the photoionization current and depended on a variable excitation energy. For a given excitation energy E_a the photoionization current is proportional to the integral over the PE spectrum from $E_B = 0$ to $E_B = E_a$. Even if a sufficiently variable source is available for excitation (which was not the case prior to the invention of synchrotron radiation), an energy-resolved method is always preferable to an integral one. In an energy-resolved measurement, only electrons with a kinetic energy that falls in the window defined by the resolution of the analyzer contribute to the statistical noise, whereas in an integral method, all electrons with $E_{kin} \le E_a$ contribute. The older methods therefore permitted determination of only the first or, under fortunate conditions, the first few binding energies. The photoemission processes leading to higher excited final states only became accessible with PES. In

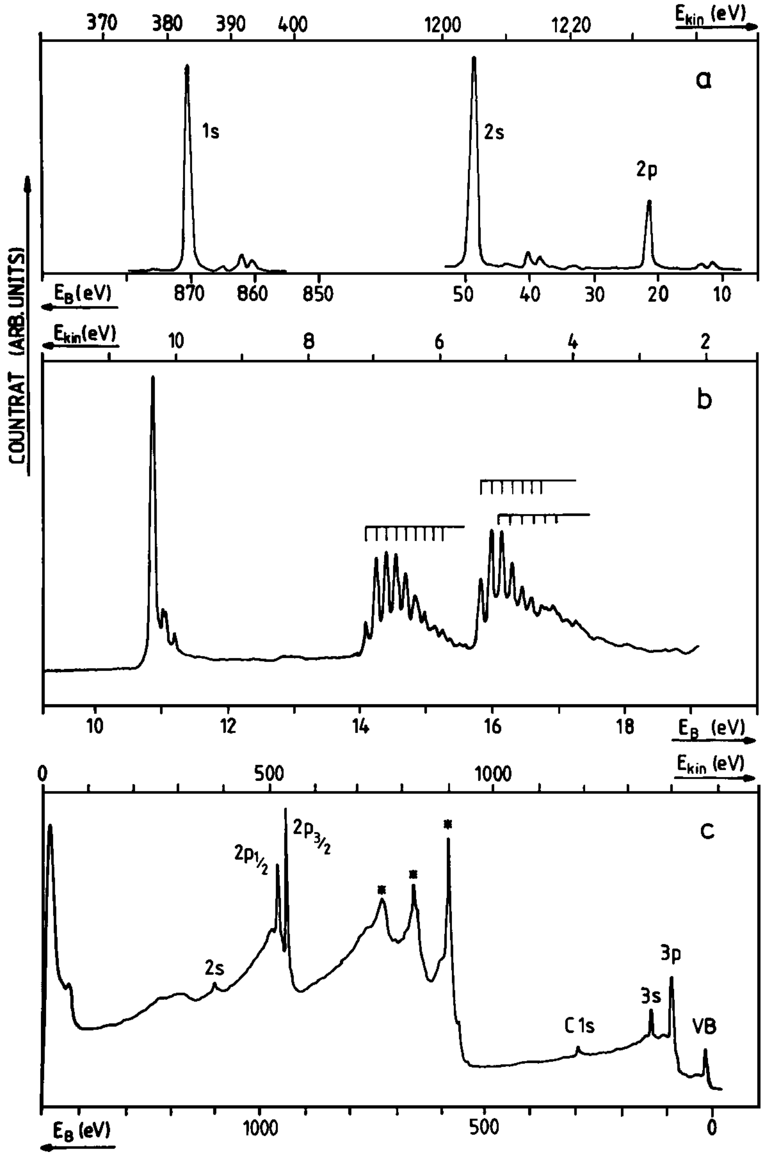


FIGURE 1 Examples of photoelectron spectra: (a) neon, $h\nu = 1253.6$ eV; (b) formaldehyde, $h\nu = 21.2$ eV; (c) metallic copper, $h\nu = 1486.7$ eV. [Spectrum of neon, from Siegbahn, K., *et al.* (1971). "ESCA Applied to Free Molecules," Elsevier, Amsterdam.]

Section II we will discuss in detail why and how these processes in particular contribute to a better understanding of the electronic structure of the systems investigated. First, however, we will clarify two terms frequently used in connection with PES: low- and high-energy PES. We will also discuss some historical aspects that will aid the reader in understanding the development of PES.

In low-energy PES, usually termed ultraviolet PES or UPS, far-UV radiation is used for excitation. The most common source for UPS is the helium resonance lamp, which provides radiation with an energy of 21.2 eV and a half-width down to 10 meV (see Section III for further details). Only valence electrons can be photoionized with this energy. The development of UPS in its application to free molecules is mainly due to work by D. W. Turner and associates in Oxford, who reached a breakthrough in the early 1960s.

In high-energy PES, characteristic x-irradiation with energies between 100 and 2000 eV and half-widths of about 1 eV is used for excitation. Because of the higher excitation energy, valence electrons as well as electrons from inner shells (core electrons) can be photoionized. The core electrons yield direct information on elemental composition and on the chemical state of a given element. Changes

in the chemical state, like different oxidation states or exchange of substituents, lead to small changes in the binding energy of core electrons. These changes or "chemical shifts" are the key to most analytical applications of high-energy PES. Because of the lower resolution, the information obtained for valence electrons is more limited than in UPS. The development of high-energy PES, which is usually termed *electron spectroscopy for chemical analysis* (ESCA) or *X-ray photoelectron spectroscopy* (XPS), began in 1914 with Robinson and Rawlinson's investigation of the photoemission of gold. Steinhardt and Serfass succeeded in 1951 in the first analytical application, which was then perfected by K. Siegbahn and coworkers in Uppsala in the late 1950s. In 1981 Siegbahn was honored with the Nobel Prize for this significant development. Much of his pioneering work on the analysis of solid organic and inorganic materials was published in a single volume in 1967. A second volume following 2 years later contained applications to free molecules.

C. Competing Photoemission Processes

The photoemission processes $M_0 \to M_i^+$ are not the only processes that lead to the appearance of electrons with well-defined kinetic energies. Depending on the energy of the exciting radiation, some other processes also contribute to the PE spectrum. These processes are indicated by the circled numbers in Fig. 2, which shows a schematic diagram for the electronic states of the initial system (M) and the systems that have lost one (M^+) or two (M^{2+}) electrons.

Process 1 is the process discussed in the previous section. It corresponds to photoionization by emission of either a valence (1a) or a core (1b) electron.

Process 2 represents an autoionization. If the energy $h\nu$ of the exciting radiation coincides with an electronic transition of the neutral system (which in the valence region is very likely for larger molecules), a photon can be absorbed by the neutral system. In case the final state M_k of this process has a higher energy than the ground state of M^+, M_k can decompose into an electron and a low-lying state of M^+. Since both the initial and the final state of the autoionization (AI) process are well-defined electronic states, the electron created in this process has the well-defined kinetic energy

$$E_{kin}^{AI}(e^-) = E(M_k) - E(M_i^+) \tag{3}$$

In Eq. (3) we again neglected the rebound energy of the heavy particle. Unlike the kinetic energy of an electron produced in process 1, $E_{kin}^{AI}(e^-)$ does not depend on $h\nu$. Therefore, the binding energy scale has no meaning for autoionization peaks. These peaks appear primarily at low kinetic energies since autoionization is usually efficient only when the initial and final states of the process are close in energy.

Process 3 in Fig. 2, is a direct transition from M_0 to states of M^{2+} under simultaneous emission of two electrons. As discussed in Section II, such a process has a much lower probability than process 1 and leads to a continuous energy distribution of the produced photoelectrons.

Process 4 represents an Auger transition. As discussed above, photoionization can lead to the creation of a core hole provided the energy of the exciting radiation is high enough. The resulting state of M^+ is highly excited. In about 10^{-16} sec, it relaxes to a lower excited state of M^+ by emission of an X-ray photon or to a lower lying state of M^{2+} by emission of another electron. The latter transition, known as an Auger process, has a higher probability for light atoms, up to about $Z = 40$. Since the initial and final states of an Auger transition are well-defined electronic states of M^+ and M^{2+}, respectively, the emitted Auger electron has the well-defined kinetic energy

$$E_{kin}^{Au}(e^-) = E\left(M_j^+\right) - E\left(M_k^{2+}\right) \tag{4}$$

Auger transitions contribute strongly to high-energy PE spectra. For example, in Fig. 1c all peaks indicated with an asterisk result from Auger transitions. As in autoionization, the kinetic energy of the Auger electron does not depend on $h\nu$. Autoionization and Auger processes, therefore, can be separated from photoemission processes by variation of the excitation energy.

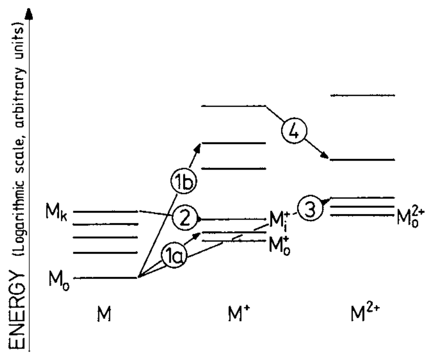


FIGURE 2 Schematic representation of different processes that lead to the appearance of electrons with well-defined kinetic energy.

D. Comparison to the Optical Spectrum of the Ion

From inspection of Fig. 2, it is clear that the information obtained from a PE spectrum of the system M is basically the same as the information obtained from an optical spectrum of M^+. The difference between the binding energy $E_B(i)$ corresponding to the photoemission process

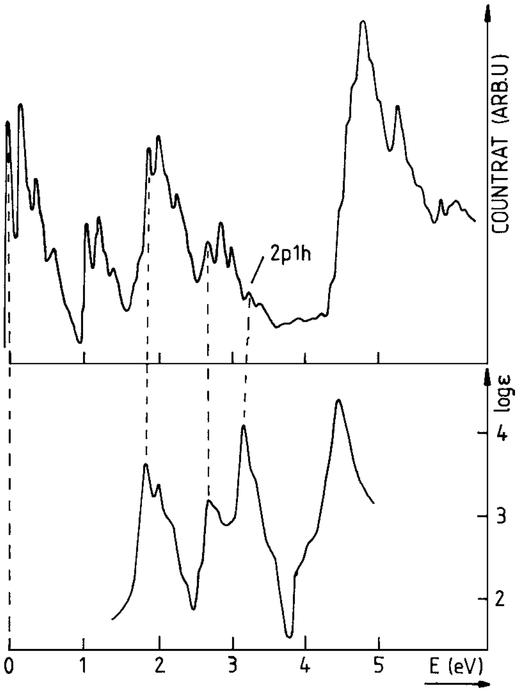


FIGURE 3 Comparison of the PE spectrum of octafluoronaphthalene to the absorption spectrum of octafluoronaphthalene cation.

$M_0 \to M_i^+$ and the binding energy $E_B(0)$ corresponding to $M_0 \to M_0^+$ is equivalent to an optical transition between M_0^+ and M_i^+. If the 0–0 transition of the first photoemission band is used as the origin of a new energy scale, the energy of the higher lying ionization bands (or lines) corresponds to an excitation energy within the system M^+, which in principle can be obtained by optical spectroscopy. This is shown in Fig. 3, where the upper and lower spectra are the PE spectrum of octafluoronaphthalene and the absorption spectrum of the octafluoronaphthalene cation, respectively. The two spectra are arranged so that the origin of the energy scale of the absorption spectrum is matched to the first peak of the PE spectrum. Considering that the PE spectrum was taken in the gas phase and the absorption spectrum in solution, the similarity of the spectra is striking. Most of the bands that are seen in the optical spectrum are also seen in the PE spectrum and vice versa. However, the intensities of the bands are different (note that the absorption spectrum has a logarithmic intensity scale).

In spite of the similarities between the PE spectrum of a molecule and the optical spectrum of the corresponding cation, the spectra often yield different types of information:

1. The connection described above holds only for molecules. For extended systems like solids it is no longer valid.

2. For molecules it is often much more difficult to measure the optical spectrum of an ion than to measure the PE spectrum of the parent molecule.

3. States with an excitation energy higher than about 6 eV are difficult to study in optical spectroscopy. This is particularly important for highly excited ion states that correspond to the removal of an inner shell electron.

4. For reasons discussed in Section II, the selection rules are quite different for both types of spectroscopy. The number of transitions allowed in photoemission is usually much smaller than the number of allowed optical transitions, especially at high excitation energies. Therefore, PE spectra are considerably easier to interpret than optical spectra, as will be seen in the following sections.

E. Photoelectron Spectra of Solids

In the case of solids some special aspects, such as the very limited escape depth, must be considered. A photoelectron created inside a solid must escape into the vacuum to be measured. Since the probability of inelastic scattering is very high as long as the electron moves inside the solid, only electrons created close to the surface have a chance to escape without a secondary energy loss. Figure 4 shows the average escape depth as a function of the kinetic energy of the electron together with some of the most useful excitation lines (see Section III). For a kinetic energy of about 100 eV, the escape depth is lowest,

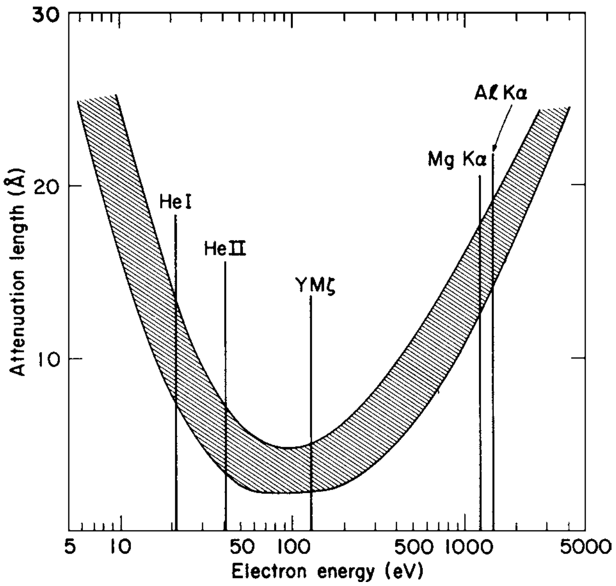


FIGURE 4 Connection between escape depth (attenuation length) and kinetic energy of the photoelectron. [From Cardona, M., and Ley, L., eds. (1978). "Topics in Applied Physics," Vol. 26, "Photoemission in Solids I," p. 193, Springer-Verlag, Berlin.]

with an average value of only a few angstroms. Even for a kinetic energy of 10 or 1000 eV, the escape depth is only on the order of 20 Å. PES probes only the few outermost atomic layers of a solid, which can be a disadvantage if one wants to study the bulk material. First, the composition of the surface is often different from the composition of the bulk because of segregation effects or surface contamination (Section III.A). Even if there is no difference in composition, there is usually a strong contribution from the outermost layer, especially for kinetic energies around 100 eV. The outermost layer is chemically always different from the interior, since the atoms in this layer have fewer neighbors. The surface sensitivity of PES is advantageous, however, if we want to study the surface itself. Therefore, PES has become one of the most powerful tools in surface science. It allows us to study not only a surface, but also atoms or molecules sitting at the surface. Nowadays it is possible to detect coverages down to a fraction of a monolayer. Thus, PES is extremely useful for the investigation of adsorbates.

Another special aspect of solids is the "reference problem." For an atom or molecule in the gas phase, ionization leads to the creation of an electron and a positive ion. The electron is either detected or lost at the walls of the instrument. The ion also leaves the ionization region rapidly. By calibration with accurately known binding energies (see Table I), the binding energies of the sample can be referred to the vacuum level that corresponds to an infinite separation of electron and ion.

TABLE I Useful Calibration Lines

Atom	Level	Compound/phase	Energy (eV)[a]
Ne	1s	Gas	870.37
F	1s	CF_4/gas	695.52
O	1s	CO_2/gas	541.28
N	1s	N_2/gas	409.93
C	1s	CO_2/gas	297.69
Ar	$2p_{3/2}$	Gas	248.62
Kr	$3p_{3/2}$	Gas	214.55
Kr	$3d_{5/2}$	Gas	93.80
Ne	2s	Gas	48.47
Ne	2p	Gas	21.59
Ar	3p	Gas	15.81
Cu	$2p_{3/2}$	Metal	932.8
Ag	$3p_{3/2}$	Metal	573.0
Ag	$3d_{5/2}$	Metal	368.2
Cu	3s	Metal	122.9
Au	$4f_{7/2}$	Metal	83.8
Pt	$4f_{7/2}$	Metal	71.0

[a] For the metals the energies refer to the Fermi level instead of the vacuum level.

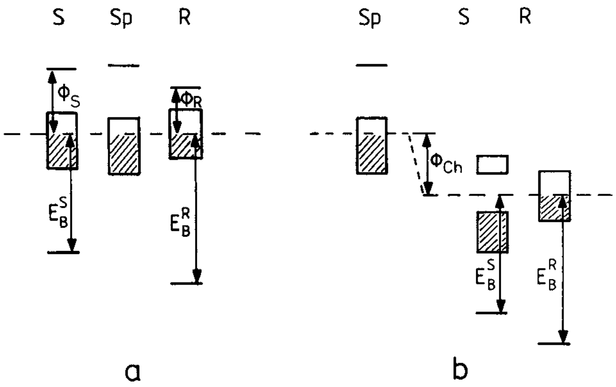


FIGURE 5 Reference schemes for solid samples: (a) conducting sample; (b) insulating sample with reference material on top. S, Sample; Sp, spectrometer; R, reference.

In the case of solids it is necessary to distinguish between conductors and insulators. If the sample is a conductor and in electrical contact with the spectrometer, the Fermi levels E_F equilibrate (Fig. 5a). The same is true for any metal that is used to calibrate the binding energy scale. The binding energy E_B^S of an arbitrary conducting sample can therefore be referred to the Fermi level of the spectrometer, which is the reference level used in most investigations. To refer E_B^S to the vacuum level of the sample, the work functions of the reference material and sample must be known.

The situation is more difficult for insulating samples. Photoionization creates positive charges within the sample that are not equilibrated immediately, and the sample becomes charged. At the same time there is usually a relatively high density of low-energy electrons close to the sample surface, which can neutralize the positive charges. The equilibrium between outgoing and incoming electrons depends on the measuring conditions, specifically on the intensity of the ionizing radiation and the cleanness of the surrounding metal parts. Therefore, the actual charging potential ϕ_{ch} (Fig. 5b) depends on the measuring conditions. The charging may not even be homogeneous over the surface area investigated (differential charging), resulting in a broadening of the observed lines. Sample charging can be reduced by use of very thin samples or a separate source of low-energy electrons (flood gun). Alternatively, sample charging can be taken into account by depositing small amounts of a reference material (usually gold) onto the sample surface or by using the carbon that is found on nearly every surface as a reference material. Assuming that the reference material and the sample are at the same potential in the irradiated area (Fig. 5b), the binding energies for the sample are then referred to the binding energies of the reference material.

F. Angular Distribution of Photoelectrons

For a single atom or molecule, the probability of emission of an electron into a certain direction with respect to an internal coordinate system is not isotropic. It depends on the initial and final states of the photoemission process, the orientation of the electric vector **E** of the ionizing radiation, and the energy $h\nu$. For example, if the electron is removed from an s orbital of an atom, the probability of finding the outgoing electron under an angle ϕ with respect to **E** is proportional to $\cos^2\phi$. Thus, the probability distribution looks like an atomic p orbital.

In a gaseous sample the molecules are randomly oriented with respect to a laboratory fixed coordinate system. To derive the angular distribution $I(\phi)$ in the laboratory system, we must integrate over all possible orientations of the molecules. If the ionizing radiation is linearly polarized, $I(\phi)$ can be expressed as

$$I(\phi) = (\sigma/4\pi)[1 + (\beta/2)(3\cos^2\phi - 1)] \qquad (5)$$

where σ is the isotropic cross section and β the "asymmetry parameter." The possible range of β is -1 to $+2$. Both σ and β depend on the initial and final states of the photoemission process as well as the kinetic energy of the outgoing electron, with σ often showing strong variations at low kinetic energies and a smooth decrease at high kinetic energies (Fig. 6). At low kinetic energies, pronounced maxima in σ, called "shape resonances," are often observed. They result from an interaction with quasi-bound states lying in the ionization continuum or from autoionization channels.

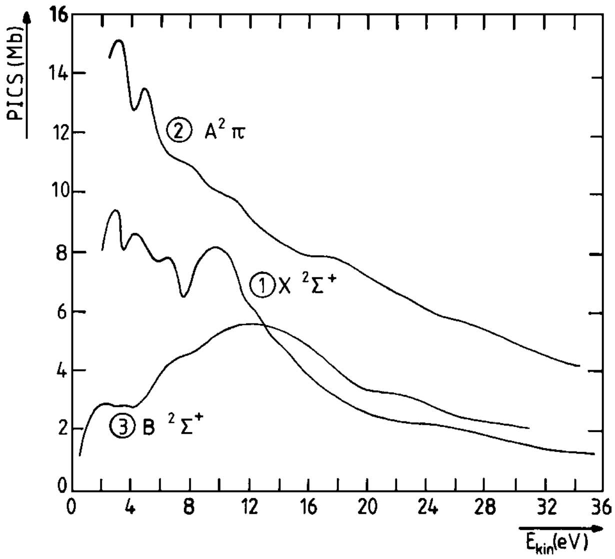


FIGURE 6 Photoionization cross section for the first three ion states of CO. [From Plummer, E. W., *et al.* (1977). *Phys. Rev. A* **15,** 2339.]

For unpolarized exciting radiation, the emission is isotropic around the incident beam. In this case the angular distribution depends only on the angle θ between the propagation direction of the radiation and the direction of the outgoing electron. We obtain

$$I(\theta) = (\sigma/4\pi)[1 + (\beta/2)(3/2\sin^2\theta - 1)] \qquad (6)$$

It is possible to measure σ and β even with unpolarized radiation. If the emitted photoelectrons are observed under an angle of $\theta = 54°44'$ (the "magic angle"), the term $(3/2\sin^2\theta - 1)$ vanishes and the isotropic cross section σ is obtained directly. This "magic angle" is used in most commercially available PE spectrometers, also in those which are designed for the investigation of solid samples.

G. Angle-Resolved Photoelectron Spectra

A different situation occurs if a molecule is adsorbed at a surface and thereby fixed in space. For example, consider a rodlike molecule (e.g., CO) which, for a given final state M_i^+, emits electrons preferentially in the direction of the molecular axis. In addition, assume that the emission probability is proportional to $\cos^2$ of the angle between molecular axis and electric vector **E** of the ionizing radiation. The system is fully described by three angles and the polarization of the radiation. The three angles are shown in Fig. 7a: Ω is the angle between the surface normal **n** and the propagation direction **s** of the radiation, θ is the angle between **n** and the direction **D** of the outgoing electron, and ϕ is the angle between the **ns** plane and the **nD** plane. Usually Ω is called the "incidence angle," θ the "polar angle," and ϕ the "azimuthal angle."

If the molecule is standing on the surface (Fig. 7b) and the radiation is polarized perpendicular to the **ns** plane (p-polarization), there is no component of the electric vector in the direction of the molecular axis and no emission at all. If the radiation is polarized in the **ns** plane (s-polarization), there is emission in the direction of the surface normal which is strongest for grazing incidence (Ω close to 90°) and vanishes for normal incidence ($\Omega \approx 0°$). Now consider the situation where the molecule is lying flat on the surface with the long axis perpendicular to the **ns** plane (Fig. 7c). For p-polarized radiation the emission is strongest for $\phi = 90°$ and large polar angles (θ close to 90°) but there is almost no dependence on Ω. For molecules lying flat but randomly oriented on the surface and s-polarized radiation, there is still no emission in the direction of the surface normal. It is therefore possible to decide from angle-resolved PES (ARPES) whether a molecule is standing or lying on a surface.

The example above illustrates how ARPES can contribute to the study of adsorbate systems. It also shows the importance of the use of polarized radiation. Another

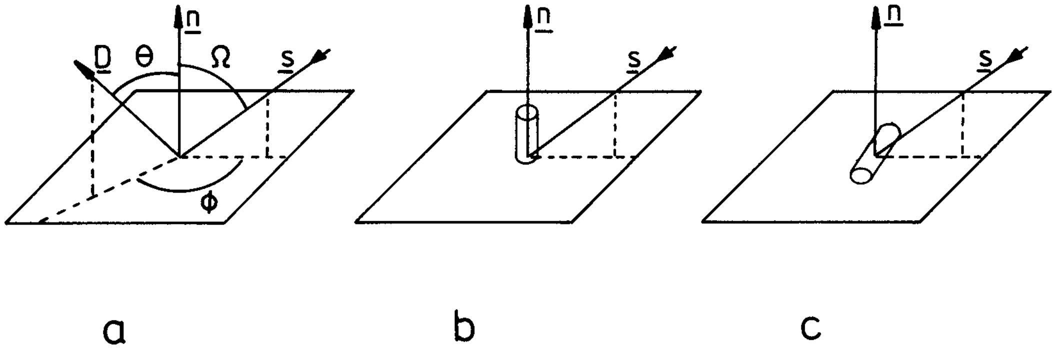


FIGURE 7 Angle-resolved photoelectron spectroscopy (see text for details).

field in which ARPES is important is the investigation of two- and three-dimensional periodic structures. Two-dimensional periodic structures are found in layer compounds or in well-ordered adsorbate systems formed at the surface of a single crystal. Three-dimensional periodic structures are found in all crystalline materials. In most cases the application of ARPES asks for single crystals of at least a few millimeters in the two surface dimensions. With an illuminated area of less than about 1 mm^2 one either loses too much intensity or risks too much radiation damage.

The important difference between free or space-fixed single molecules and periodic structures is as follows. In the first case, the intensity of the peaks in the PE spectrum is angle dependent but the position of the peaks is not. In the second case, the energy of the most prominent features in the spectrum of the valence region depends on the angles Ω, θ, and ϕ as well as the polarization of the ionizing radiation. One therefore speaks of energy distribution curves instead of PE spectra in such a case. We will come back to this type of investigation in Section IV.B.

ARPES has become more and more common in the investigation of substrate/overlayer structures. An overlayer is a zone of thickness d that differs in chemical composition from the substrate as, for example, a metal oxide on the surface of a metal. In materials science, the knowledge of d is often very important (see, e.g., Fig. 30 in Section IV.E). To determine d, one makes use of the limited escape depth of photoelectrons and the difference in binding energy that leads to distinguishable peaks for pure metal and metal oxide. The larger the d, the smaller becomes the intensity ratio $I_{\text{metal}}/I_{\text{metal oxide}}$. Increase of the polar angle θ increases the effective thickness of the overlayer and leads to characteristic changes in the intensity ratio. Equations that connect the observed θ dependence of the intensity ratio with d were derived under the following two assumptions: (1) the surface is atomically flat and (2) electrons which reach the detector without energy loss have traveled in straight lines (no elastic scattering). In the beginning, it was assumed that it is the inelastic mean free path (IMFP) that has to be used in connection with these equations. The IMFP is an intrinsic property of a material which can be measured using other techniques, such as electron backscattering. In the mid-1980s it was found that the appropriate length is not the IMFP but the attenuation length (AL). Powell defined the AL as "a value resulting from overlayer-film experiments on the basis of a model in which elastic electron scattering is assumed to be insignificant." The AL is typically 10 to 25% shorter than the IMFP because of the contribution of elastic scattering to the IMFP. If the thickness of the overlayer is larger than about three times the AL, the substrate is no longer detectable. This limits the applicability of ARPES to overlayers that are not more than a few nanometers thick. Some terms concerning electron transport in matter are defined in Table II.

To convert the "composition versus take-off angle" curves obtained from an angle-resolved XPS (ARXPS)

TABLE II Some Definitions with Regard to Electron Transport in Matter

IMFP	*I*nelastic *M*ean *F*ree *P*ath. The average distance that an electron with a given energy travels between inelastic collisions.
ED	*E*scape *D*epth. The distance normal to the surface at which the probability of an electron escaping without significant energy loss due to inelastic processes drops to e^{-1} (38%) of its original value.
AL	*A*ttenuation *L*ength. The average distance that an electron with a given energy travels between inelastic collisions as derived from a particular model in which elastic scattering is assumed to be negligible.
ID	*I*nformation *D*epth. The average distance normal to the surface from which a specified percentage of the detected electrons originates.
Sampling Depth	Three times the IMFP corresponding to an ID for which the percentage of detected electrons is 95%.

experiment to a plot of "depth versus composition," it is necessary to deconvolute the observed data. A number of algorithms for such a deconvolution has been published but as the experimental data always contain some amount of statistical error, it is often more appropriate to compare curves obtained for certain layer models with the experimental data.

H. Spin-Resolved Photoelectron Spectra

In spin-resolved photoelectron spectroscopy (SRPES), one makes use of the fact that a single electron has a spin. In a properly designed instrument, the orientation of the spin with respect to the laboratory frame does not change from the moment the electron leaves the sample to the moment it hits the detector, or if it changes, it changes in a well-defined manner. If the detector is sensitive to the orientation of the spin, as for example, a Mott-detector (see Section III.E), information is obtained on the spin distribution of the photoelectrons that leave the sample. For a nonmagnetic sample, the spin distribution is usually random, and no additional information is gained. For a magnetic sample, however, the ratio of the number of electrons that are emitted with spin-up to the number of electrons with spin-down depends on the band structure of the system and therefore on the binding energy. The combination of spin-resolved and angular-resolved PES has contributed tremendously to the understanding of all kinds of magnetic materials.

I. Zero Kinetic Energy Photoelectron Spectra

Over the last years *ze*ro *k*inetic *e*nergy (ZEKE) spectroscopy has become the favorite method for studying the vibrational and rotational structure of ions. If a laser is tuned to the exact energy difference between a given ion state and the molecular ground state, the produced photoelectrons have no kinetic energy (Eq. 2). Measuring these electrons as a function of laser frequency would lead to a PE spectrum (ZEKE-spectrum) the resolution of which depends only on the bandwidth of the exciting laser. The problem is that the exciting laser also produces photoelectrons with kinetic energy from lower lying ion states and that it is impossible to separate the ZEKE electrons from these other electrons with sufficient resolution. This difficulty can be overcome by making use of the fact that highly excited Rydberg states ($n \approx 200$) which lie a few wavenumbers below each ion state have lifetimes of the order of micro- to milliseconds. Exciting these high-n Rydberg states with a nanosecond laser pulse and waiting for a few microseconds lets all photoelectrons with nonzero kinetic energy disappear from the excitation region. An electrical pulse then field ionizes the Rydberg states. The resulting electrons are ejected from the excitation region and measured. This combination of laser excitation and pulsed-field ionization makes it possible to resolve the rotational structure of small molecular ions and the vibrational structure of larger ones.

II. PHOTOEMISSION FROM ATOMS, MOLECULES, AND SOLIDS

A. The Independent-Particle Model

To elucidate the information obtainable by PES, we start with the independent-particle model. This model is best known for atoms. The electronic structure of an atom is described by orbitals $\varphi(q_i)$, which are functions of the spatial coordinates q_i. Each orbital can hold a maximum of two electrons if these electrons have antiparallel spin. With each orbital we can associate an orbital energy ε_i. According to the aufbau principle, the state of lowest total energy is reached if the orbitals are filled in the order of increasing orbital energy. Some of the orbitals are degenerate with respect to orbital energy, and Hund's rule states that degenerate orbitals are first filled singly with electrons of parallel spin. The orbitals holding the inner shell or core electrons are energetically well separated from the orbitals of the outermost or "valence" electrons (Table III). For most atoms the valence electrons have binding energies of less than 40 eV. If the state of lowest energy consists only of doubly occupied orbitals, it is called a "closed shell ground state." This situation is depicted in Fig. 8a. The rest of this article will be restricted to the discussion

TABLE III Calculated Orbital Energies (ε_i), Experimental Binding Energies (E_B), and Relaxation Energies (R_i) for Formaldehyde and Water

Molecule	No.	$-\varepsilon_i^a$	E_B	R_i	Symmetry	Type
$H_2C{=}O$	8	12.04	10.80	1.24	$2b_1$	n(O)
	7	14.69	14.5	0.2	$1b_2$	π(CO)
	6	17.80	16.0	1.8	$5a_1$	σ(CH)
	5	18.91	16.6	2.3	$1b_1$	σ(CH)
	4	23.65	20.5	3.1	$4a_1$	σ(CO)
	3	38.39			$3a_1$	O 2s
	2	308.63	294.47	14.16	$2a_1$	C 1s
	1	559.97	539.44	20.53	$1a_1$	O 1s
H_2O	5	13.72	12.6	1.1	$1b_2$	n(O)
	4	15.70	14.7	1.0	$1b_1$	σ(OH)
	3	19.25	18.4	0.8	$3a_1$	σ(OH)
	2	36.56	32.2	4.4	$2a_1$	O 2s
	1	559.33	539.7	19.6	$1a_1$	O 1s

[a] HF-result with cc-pVTZ basis set based on DFT-B3LYP geometries obtained with the same basis set.

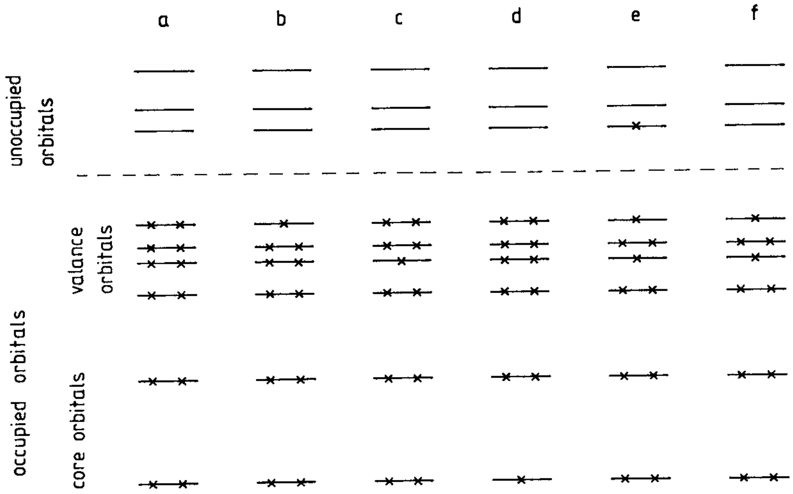


FIGURE 8 Orbital representation of (a) the initial and (b–f) different final states of the photoemission process.

of systems with such a ground state. It is an exception for atoms but is very common for molecules. Apart from the orbitals that are occupied in the ground state, there are unoccupied or "virtual" orbitals. For example, for the neon atom, with a ground state configuration $1s^2 2s^2 2p^6$, the lowest unoccupied orbital is $3s$. In the following discussion, it is important to keep in mind that "occupied" always refers to the set of orbitals occupied in the ground state M_0. Correspondingly, "unoccupied" always refers to the set of orbitals not occupied in the ground state.

When we turn from atoms to molecules, the situation changes only slightly. The major difference is that the valence orbitals no longer belong to a single atom. To a greater or lesser degree, they are extended over the whole molecule. The same is true for the unoccupied orbitals. The core orbitals are still localized at their respective atoms. Degenerate valence orbitals are less common in molecules than in atoms; they are usually found in molecules of higher symmetry with at least one three-fold axis.

For solids, the number of atoms and therefore the number of orbitals goes to infinity. The valence orbitals develop into continuous energy bands. The details of the electronic structure of a crystalline solid are described in terms of the "band structure." The energy up to which the bands are occupied is the Fermi energy (E_F) already mentioned in Section I.E. When the Fermi energy lies within a band, the solid is a metal; when it lies in a gap between two bands, the solid is a semiconductor (small gap) or an insulator (large gap). As in molecules, the core orbitals in solids still behave like those in atoms.

B. The Frozen Orbital Approximation

We now return to photoemission. For a first approximation, we assume that we can remove a single electron from one of the occupied orbitals without disturbing the remaining electrons. This is called the "frozen orbital approximation" (FOA). This process creates a hole in the manifold of the occupied orbitals, and we call the resulting state a "hole state." If the electron is removed from the highest occupied orbital, the final state of the photoemission process is the ground state M_0^+ of the ion (Fig. 8b). If the electron is removed from a lower lying orbital, an excited state of the ion is reached which we call a "hole excited state" because moving a hole downward is the same as moving an electron upward (Fig. 8c). When the electron is removed from a core orbital, we speak of a "core hole state" (Fig. 8d). Such a state is usually labeled by the chemical symbol of the atom and the orbital from which the electron has been removed (e.g., C $1s$, O $1s$, P $2p$).

If we start from M_0^+ (Fig. 8b) and excite one of the remaining electrons to an unoccupied orbital (Fig. 8e), a new type of excited ion state is reached. Compared to the ground state M_0 of the initial system, this state has two holes in the occupied orbitals and one electron in an unoccupied orbital. Therefore, it is called a "two-hole one-particle (2h1p) state." Analogously, we can define three-hole two-particle states, and so on.

C. Secondary Structures in PE Spectra

The nh(n-1)p states are frequently called "shake-up states," based on the idea that photoionization is such a strong perturbation that the whole electron system is "shaken" and one or more of the remaining electrons are "shaken up" to unoccupied orbitals. If the "shake-up electron" receives enough energy to leave the system, the final state of the photoionization is a state of the doubly ionized system (Fig. 8f). This type of transition was shown as

process 3 in Fig. 2 and, in a further extension of the above ideas, is called a “shake-off” process. Since the second electron is no longer bound, shake-off excitations lead to a continuum. This corresponds to the continuum following discrete excitations in the optical spectrum of the ion.

Examples of the processes discussed above are seen in the photoemission spectrum of neon (Fig. 1a). The first line appears at 21.6 eV, corresponding to photoionization of one of the six $2p$ electrons. The second line, at 49 eV, corresponds to photoionization of a $2s$ electron and the third line, at 870 eV, to photoionization of a $1s$ electron. On an enlarged intensity scale (Fig. 9) shake-up satellites are seen at the high binding energy side of the Ne $1s$ line, which result from transitions into 2h1p states. The lines designated 3, 5, 6, and 7, for example, have been identified as transitions into states where, in addition to the creation of the $1s$ hole, a $2p$ electron is excited to $3p$, $4p$, $5p$, and so on. In addition, a shake-off continuum underlies the high-energy part of the $1s$ shake-up spectrum. Shake-up satellites are especially common for the $2p$ spectra of metals with d-electrons, as for example in the Cu $2p$ spectrum of copper(II) oxide shown in Fig. 10a, and in the spectra of aromatic organic compounds where they result from transitions between bonding and antibonding π-orbitals. Much weaker shake-up satellites have been reported for the $\sigma \rightarrow \sigma^*$ transitions in aliphatic organics.

The appearance of shake-up structures is not restricted to the core region. Because of the close spacing of the valence orbitals and the vibrational structure of the corresponding photoemission bands (Section II.G), shake-up structures in the valence region are more difficult to observe. Figure 3 shows an example where it was possible to detect a shake-up structure (indicated 2h1p) bycomparison with the optical spectrum of the ion. A transition from the ion ground state M_0^+ to this particular 2h1p state is allowed and leads to a strong band in the absorption spectrum of the ion. In the PE spectrum, which starts from M_0 instead of M_0^+, the 2h1p final state gives rise to only a weak structure. From the PE spectrum alone it would be difficult to detect this weak structure as a result of an independent transition. Comparison with the UV spectrum of the ion, however, allows an unambiguous assignment.

In the examples shown, the shake-up transitions only lead to weak structures in the PE spectrum. This is a very general observation. Only in special cases can shake-up satellites reach an intensity comparable to that of the main line (see Section IV.D and Fig. 28). Why are shake-up transitions usually weak? In the FOA, transitions are allowed only when the initial and final states differ by one electron. Transitions that involve more than one electron are forbidden. Since all 2h1p states differ from the initial state by at least two electrons (Fig. 8), transitions from M_0 to 2h1p states (and also to shake-off states) are strictly forbidden in the FOA. Only transitions to final single-hole states are allowed. When degenerate orbitals are counted only once, the number of lines or bands observable in a PE spectrum

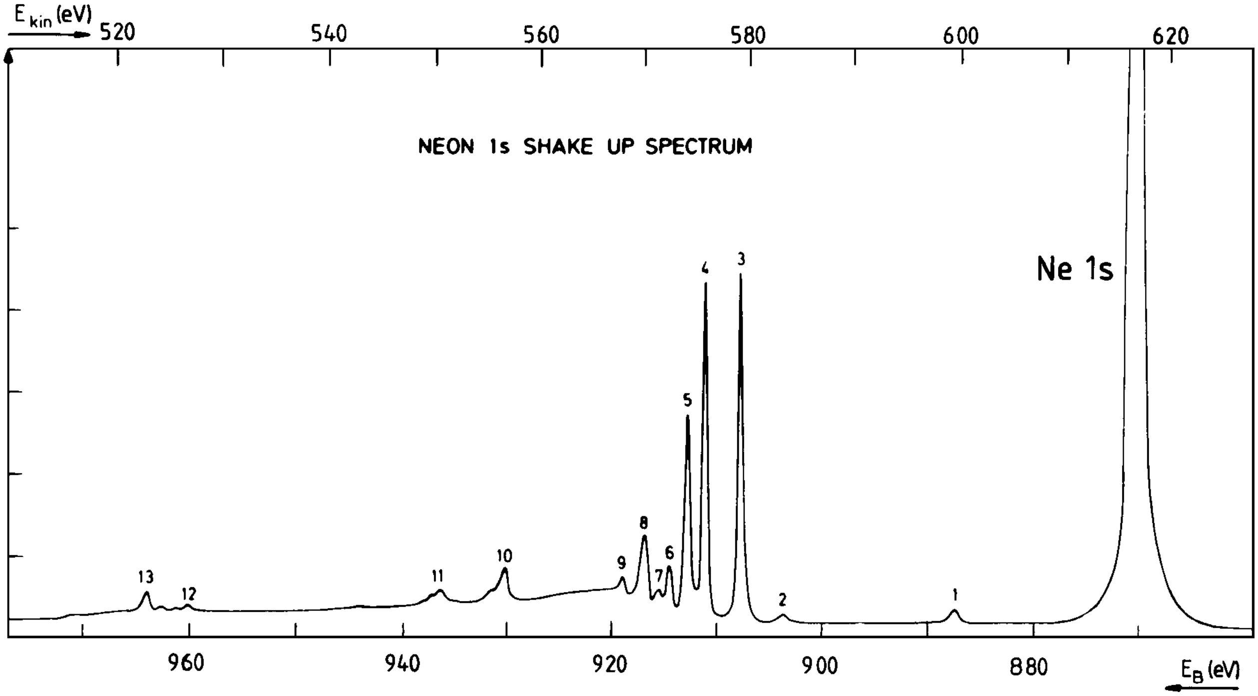


FIGURE 9 Shake-up satellites accompanying Ne $1s$ photoionization; $h\nu = 1486.7$ eV. [From Siegbahn, K. (1974). *J. Electron Spectrosc.* **5**, 34.]

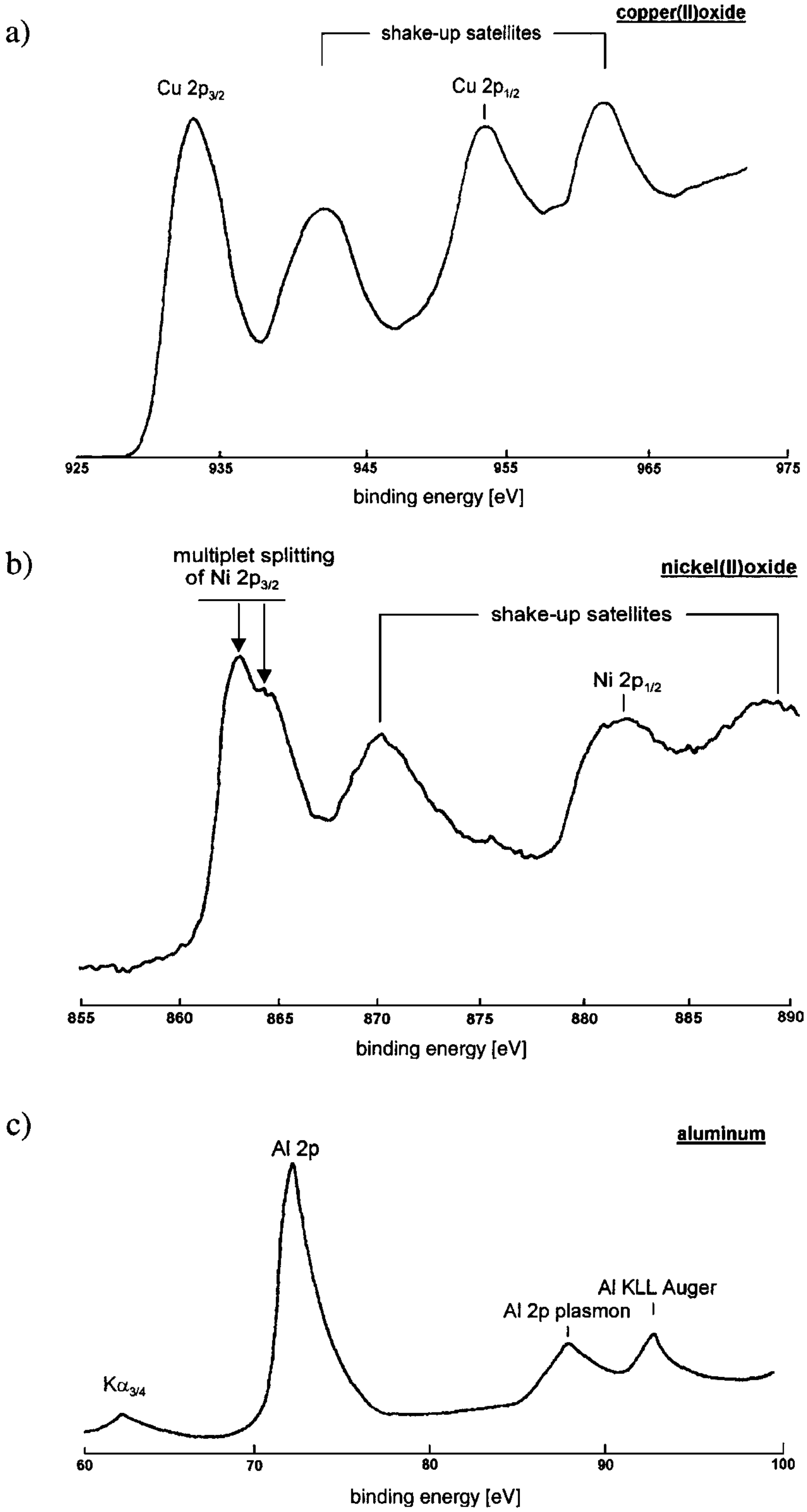


FIGURE 10 Energy loss features in XPS spectra: (a) Cu 2*p* shake-up satellites, (b) Ni 2*p* multiplet splitting, (c) plasmon loss features from clean aluminum. [From Watts, J. F. (1994). *Vacuum* **45**(6/7), 653.]

should be equal to the number of occupied orbitals. Therefore, all the prominent bands in the valence region of free molecules and all main lines in the inner shell region of free molecules and solids are usually assignable in the single-particle picture. Basically, this type of interpretation also holds for photoemission from solids. However, in this case some special aspects must be considered, which will be discussed in Section IV.B.

Additional structures in PE spectra can result from multiplet splitting (Fig. 10b shows an example), which is discussed in more detail in Section II.H, plasmon losses, and even the X-ray source itself. Plasmon losses, which are of little analytical use, are observed in the PE and Auger spectra of clean metal surfaces, as shown in Fig. 10c for a clean aluminum surface. They result from collective oscillations in the conduction band excited by the outgoing electron which thereby suffers a discrete energy loss. The plasmon frequency and multiples thereof are observed as a characteristic series of peaks on the high binding energy side of the main core line.

The radiation source also gives rise to weak additional photoelectron signals, as, for example, both the AlK_α and the MgK_α lines have satellites some 10 eV below the main line with around 10% of its intensity. In addition to these satellites, an oxidized or damaged anode (CuK_α) as well as cross-talk from the complementary side of a badly aligned twin anode can lead to the appearance of unwanted radiation.

D. Koopmans' Theorem

The correspondence between orbital picture and PE spectrum goes even further than the one-to-one correspondence between main lines and occupied orbitals. In the FOA, the binding energies relative to the vacuum level are directly connected with the orbital energy of the occupied orbitals

$$E_B(i) = -\varepsilon_i \tag{7}$$

This relation was derived by T. Koopmans in 1932 and is now known as Koopmans' theorem (KT). It is the basis of most applications of PES in electronic structure elucidation. If KT were strictly valid, we could experimentally observe orbital or single-particle energies, which in reality exist only in the theoretical framework of the independent-particle model. It must be clearly understood that orbital energies are not observable in the sense of quantum mechanics. However, they can be calculated by a variety of different methods, and these calculations can be performed with a high degree of accuracy for small and medium size molecules. A comparison of orbital energies from such calculations with experimentally observed binding energies shows that the deviation from KT is usually less than 10% (see Table III). Therefore, KT is a close approximation.

However, one point always must be kept in mind if KT is applied: orbital energies are quantities solely related to the initial state of the photoemission process. The application of KT is equivalent to neglecting any influence of the final state on the binding energies. Obviously, this cannot be true. To take final state effects into account, we must go beyond the FOA. We return to this point in the following section.

In the application of PES to the study of molecular electronic structure, KT is seldom used in the form of Eq. (7). In most cases, a direct comparison of calculated ε_i and measured $E_B(i)$ is not the main objective. Often it is more interesting to investigate how a certain ε and correspondingly a certain E_B vary with alterations in chemical constitution. In this case we use a "weaker" form of KT:

$$\Delta E_B(i) = \Delta\varepsilon_i \tag{8}$$

where we connect binding energy shifts (ΔE_B) with orbital energy shifts. In considering these chemical shifts, we do not completely neglect final state effects. We only assume that they are approximately constant within a certain class of compounds. The chemically appealing feature of this approach lies in the great variety of models available to estimate orbital energy shifts. For example, all the models that have been developed to describe the influence of different substituents on physical properties and chemical reactivities of molecules can be applied, and frequently a more direct proof of a given model is possible with PES. Examples of this type of application are discussed in Sections IV.A and IV.C.

Of special importance is a model which connects core electron binding energy shifts to atomic charges. Chemists usually attribute partial charges q_A to the different atoms of a molecule, even though this concept is problematic from a strictly theoretical point of view. In the "point charge approximation," the energy $\varepsilon_i(A)$ of a core orbital at atom A can be expressed in terms of partial charges by

$$\varepsilon_i(A) = k(i, A)q_A + V(q_B) + k_0(i, A) \tag{9}$$

where q_A is the charge at the considered atom, $V(q_B)$ the "off-atom potential" created by the charges at all other atoms, and k_0 and k are parameters specific for atom A and orbital i. Within the limits of applicability of Eq. (8), the model allows one to estimate changes in atomic charges from measured shifts of core electron binding energies.

E. Final State Effects

In the preceding section we saw that the most prominent structures in PE spectra can be explained adequately in the

single-particle approximation. Two facts, however, clearly show the limitations of this type of interpretation: (1) the deviations from KT [Eq. (7)] and (2) the appearance of shake-up and shake-off satellites.

From these two observations, it is obvious that we cannot completely neglect final state effects. Using KT as a first approximation, we can write

$$E_B(i) = -\varepsilon_i - R_i \tag{10}$$

where R_i is called the "relaxation energy." The meaning of Eq. (10) can be understood by the following Gedanken experiment. We separate the photoemission process into two steps. In the first step, we use the FOA and remove one electron from a given orbital i, leaving all other electrons unperturbed. To remove the electron we need the energy $-\varepsilon_i$. In the second step, we consider the reaction of the remaining electrons to the presence of the hole created in the first step. The system will "relax" to a new, energetically favorable situation. Therefore, the relaxation energy is usually positive. Only in rare cases can specific quantum mechanical effects (so-called correlation effects) lead to small negative values of R. In Table III we have compared experimental binding energies to calculated orbital energies. The relaxation energies derived from these two sets of data vary from orbital to orbital. R is much larger for core electrons than for valence electrons; however, relative to the magnitude of the binding energy, R is similar for all shells.

From our Gedanken experiment, we suggest that R_i will be connected strongly to the mobility of the electrons in the system. If the orbital from which the electron is removed is highly localized, as in a core orbital or a lone pair orbital, the most efficient stabilization of the hole state will be achieved by transfer of negative charge to the vicinity of the hole. If a direct transfer is not possible, the stabilization can be achieved only by polarization of the surroundings. From this consideration we expect larger relaxation energies for the core ionization of metals, where the electrons in the valence band move almost freely, than for the core ionization of insulators, where the polarization of the nearest neighbor atoms yields the most important contribution to the relaxation. For delocalized holes that result from the photoionization of π electrons of unsaturated molecules or valence electrons of solids, the relaxation contribution is expected to be smaller and less dependent on the individual orbital. This is in accordance with the data shown in Table III; for formaldehyde, for example, the smallest relaxation energy is found for the π orbital $1b_1$.

The considerations discussed above can also be viewed in a somewhat different manner. The ion state $(M_i^+)^*$ formed in the FOA is not a real state (eigenstate) of the investigated system. It exists only in the framework of the theoretical model. It can be described, however, by a superposition of the eigenstates M_k^+ of the ion, where each of these states contributes with a certain weight factor g_k:

$$\left(M_i^+\right)^* = \sum_k g_k M_k^+ \tag{11}$$

In the high-energy limit where the electron leaves the ionized system very quickly (sudden limit), the probability of a transition to a final state M_k^+ is equal to g_k^2. The main contribution usually comes from a state M_i^+ that has the same orbital occupation as the hypothetical frozen orbital state $(M_i^+)^*$. However, because of the presence of the hole, the orbitals of M_i^+ differ somewhat from the orbitals of the initial state M_0. Since the weight factor g_i is usually greater than 0.5, we normally can give an assignment of the main bands or main lines of a PE spectrum in terms of KT, as discussed in the previous section. The remainder of the weights is frequently distributed over a variety of final states M_k^+ including the continuum states, thus explaining shake-up satellites and the shake-off continuum. When we are able to describe M_k^+ to a good approximation by a single 2h1p state, we reach a situation as discussed in the previous section in connection with Fig. 9, where several of the observed shake-up satellites could be assigned to specific electronic excitations in the core ionized system. For larger systems, where the number of possible electronic excitations becomes very large, we will be able to identify the shake-up satellites only in cases where the remaining weight is not distributed more or less equally over a large number of final states. When g_k has a somewhat larger value for a specific final state M_k^+, we will observe a characteristic shake-up satellite even for an extended system. An example of such a situation will be discussed in Section IV.D.

We now turn to binding energy shifts. From Eq. (10) we obtain

$$\Delta E_B(i) = -\Delta\varepsilon_i - \Delta R_i \tag{12}$$

Binding energy shifts depend as much on initial state effects (via $\Delta\varepsilon_i$) as on final state effects (via ΔR_i). Often we are specifically interested in initial state effects, because we want to derive information on the electronic structure of the initial system M_0 and its dependence on variations in chemical constitution. This information, however, can be derived only if ΔR is negligibly small or if we are able to obtain independent information on ΔR. The relaxation contribution itself also contains valuable information, since it is connected with electronic relaxation processes that can take place during a chemical transformation. In a wide variety of chemical reactions the transition state is charged. The better this charge can be screened by a relaxation of the whole electronic system, the lower the energy of the transition state.

A variety of theoretical models have been developed in which relaxation is taken into account (transition state models, relaxed potential models, equivalent core models). A discussion of these models is far beyond the scope of this article. Here, we will only add some comments on methods by which it is possible to separate initial and final state effects with the use of experimentally available data. These methods are based on a combination of PE and Auger electron spectroscopy. We consider an Auger transition from an initial state with a single hole in the inner shell k to a final state with two holes in another inner shell i. This Auger transition is combined with photoionization processes that correspond to the photoemission of an electron from orbital k and from orbital i. This yields

$$\Delta\beta(i) = \Delta E_{kin}^{Au}(kii) + 2\Delta E_B(i) - \Delta E_B(k) = 2\Delta R_i - \Delta R_{ii} \quad (13)$$

where ΔR_{ii} is the relaxation contribution of the double-hole final state of the Auger transition. The parameter $\Delta\beta(i)$ is independent of the reference level. Therefore, it can be obtained for molecules in the gas phase as well as for solids. Since it is independent of the reference level, it is also independent of sample charging if the Auger kinetic energy $E_{kin}^{Au}(kii)$ and the binding energies are derived from the same measurement. When we introduce the approximation that the relaxation energy results mainly from classical electrostatic contributions, ΔR_{ii} should be four times as large as ΔR_i. We then obtain

$$\Delta\beta(i) = -2\Delta R_i \quad (14)$$

If experimental constraints allow only the measurement of one of the binding energies, we can use the cruder approximation

$$\Delta\alpha(i) = \Delta E_{kin}^{Au}(kii) + \Delta E_B(i) = -2\Delta R_i \quad (15)$$

where $\Delta\alpha$ is the Auger parameter shift, introduced by C. Wagner in 1972.

The drawback of this experimental method for disentangling initial and final state effects is that it is only applicable to core ionizations and heavy atoms. An Auger transition that does not involve valence electrons can be observed only for atoms with at least two inner shells. Thus, an experimental estimate of final state relaxation effects can be derived only for third row and higher elements. For the light elements, however, highly accurate calculations are increasingly available that allow the determination of ΔR from the comparison of experimental and theoretical data, as shown in Table III.

F. Auger Parameter and Chemical State Plot

The importance of the Auger parameter reaches far beyond the determination of relaxation energies. A popular means of displaying Auger parameter data is the *chemical state plot*, sometimes also referred to as the "Wagner Plot" after the inventor of this format. Binding energies (E_B) are plotted on the abscissa and Auger kinetic energies (E_{kin}^{Au}) on the ordinate. Diagonals represent constant α-values. Figure 11 shows an example. Different positions within such a plot indicate a different chemical environment of the considered element. This is the reason for the name chemical state plot. In case of a charged sample, the charging contributions to E_B and E_{kin}^{Au} cancel in the determination of α. Charging shifts a data point only along a diagonal of a chemical state plot.

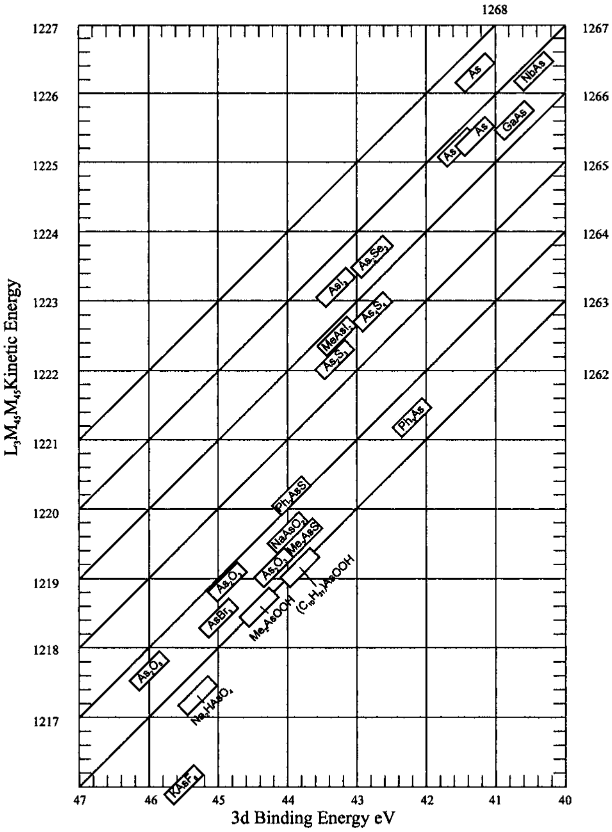


FIGURE 11 Wagner chemical state plot for arsenic. The modified Auger parameter is evaluated on the diagonal grid. [From Wagner, C. D., *et al.* (1979). "Handbook of XPS," Perkin Elmer Corp, Eden Prairie, MN.]

G. Vibrational Fine Structure

The low-energy PE spectra of small and medium size molecules often contain bands with well-resolved vibrational fine structure. Since the resolution in UPS is usually limited to about 150 cm^{-1} ($\approx$15 meV), only excitations into well-separated vibrational states of the final ion state can be observed (see, however, Section I.I). As

long as the temperature of the sample is not much higher than 300 K, vibrational excitations in the initial states do not perturb the observed spectra; low-frequency vibrations excited at these temperatures are covered by the limited resolution.

To first order, the intensity distribution of the different lines of a vibrationally structured PE band can be interpreted in terms of the Franck–Condon principle (Fig. 12). The most probable transition is the "vertical transition." In a potential diagram, where the electronic energy is drawn as a function of interatomic distances, the vertical transition is best approximated by a vertical line drawn from the minimum in the potential of the initial state to the potential curve of the final state. It is the binding energy corresponding to the vertical transition [the "vertical ionization potential" (VIP)] that must be used in connection with KT [Eq. (7)]. The transition from the lowest vibrational level of the initial state to the lowest vibrational level of the final state is called "adiabatic transition." If the minimum of the final state is not displaced with respect to the initial state, vertical and adiabatic transitions fall together (Fig. 12). The difference between adiabatic and vertical ionization potentials is therefore a measure of the change in equilibrium geometry between ion state and initial state. For example, the first ionization band of formaldehyde (Fig. 1b) shows a very intense 0-0 transition and little vibrational fine structure. From this we can conclude that the electronic ground state of the formaldehyde cation is very similar in geometry to the electronic ground state of formaldehyde. For the second photoionization band the vertical transition corresponds to the fifth vibrational level, telling us that the equilibrium geometry of the first electronically excited ion state differs considerably from the ground state of the neutral system. In the fourth ionization band of formaldehyde we observe a rapid loss of vibrational fine structure after the fifth vibrational excitation. This is a typical pattern for an excitation into a dissociative final state (see Fig. 12). Excitation with an energy greater than the dissociation limit leads to fragmentation of the ion. The vibrational structure becomes broadened in the region of the dissociation limit and then continuous. From the examples discussed it is clear that the vibrational fine structure of a photoionization band contains information on the geometry and the potential surface of the final ion state.

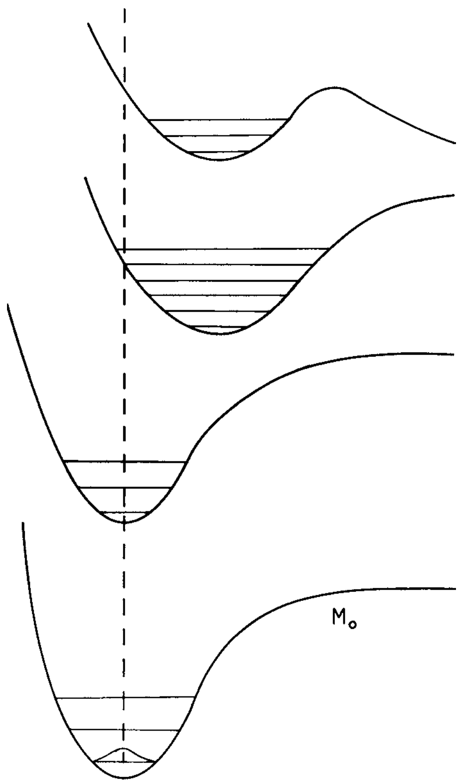


FIGURE 12 Schematic representation of potential curves for the ground state of the initial system (M_0) and for different ion states M_k^+.

For larger molecules we usually do not observe vibrationally structured bands (Figs. 19, 21, and 22). The number of possible vibrations that can be excited in the final ion state increases rapidly with the size of the system, and the superposition of the different vibrational excitations leads to more or less continuous bands. For these unstructured bands the VIP is attributed to the band maximum. Some bands show vibrational fine structure even for larger molecules (Figs. 3 and 22). In planar unsaturated compounds, for example, an ionization from the π system usually couples to skeletal modes that lie near 1400 cm^{-1}, resulting in a well-developed vibrational structure. However, for the same molecules the bands that result from ionizations from CH σ bonds are mostly broad and unstructured.

The analysis of the vibrational fine structure can be carried even further when we combine it with the assignment of the main bands in the single-particle approximation. From theoretical calculations, for example, the highest occupied orbital of formaldehyde is a nonbonding orbital (n orbital) that is mainly localized at the oxygen atom. When an electron is removed from such an orbital the geometry is expected to change only little, in agreement with what we concluded from the fine structure of the first band. A detailed analysis of this structure reveals that the CO stretching vibration, which is 1744 cm^{-1} in the ground state of the neutral molecule, is only slightly reduced to 1590 cm^{-1} in the lowest ion state. The second highest orbital is the π orbital of the CO double bond, an orbital that is strongly CO bonding. When an electron is removed from this orbital, the bond is considerably weakened. This should lead to an increased CO bond length, in accordance

with the difference between vertical and adiabatic ionization potential observed for the second band. The weakening of the CO bond is also seen in the lowering of the CO stretching frequency, which is reduced to 1210 cm^{-1} in the first excited state of the ion. These examples show how a careful analysis of the vibrational structure of photoionization bands and a comparison of observed vibrational frequencies to the vibrational frequencies known for the molecular ground state provide information on the bonding characteristics of molecular orbitals.

H. Multiplet Splitting

As mentioned earlier, core electrons behave in an atom-like way, even when the respective atom is part of a molecule or a solid. As a consequence, the angular quantum number is always a good quantum number for inner shell electrons. We can denote such an electron by its principal quantum number n and its angular quantum number l. We therefore speak of $2s$, $2p$, or $3d$ electrons and correspondingly of a $2s$, $2p$, or $3d$ hole. However, an additional effect must be taken into account for core holes: the spin–orbit coupling. Since we can treat a hole in an otherwise completely filled shell in the same way as a single electron in an otherwise empty shell, we can attribute a spin $+^1/_2$ or $-^1/_2$ to the hole. Under the action of spin–orbit coupling, the angular momentum resulting from the spin couples to the angular momentum corresponding to the spatial motion of the hole to form a total angular momentum J. The two possible J values that can be formed are

$$J = l \pm {}^1/_2 \tag{16}$$

Only $J = +^1/_2$ is possible for $l = 0$ (s orbitals). Photoionization of a core electron with angular quantum number $l > 0$ therefore leads to two possible final states that correspond to the two possible J values. Because of spin–orbit coupling, the two final states have different energies. For different inner shells, spin–orbit coupling decreases with increasing n and within the same shell it decreases with increasing l. To give an impression of the magnitude of spin–orbit splitting, data for xenon are provided as an example in Table IV. If the splitting exceeds the experimental resolution, we observe a doublet in the XPE spectrum. In the case of resolved doublets the individual lines are labeled nl_J. Table V compares this type of labeling with the notation commonly used in X-ray spectroscopy.

TABLE IV Spin-Orbit Splitting in Xenon

	Splitting (eV)	
n	$l = 1$	$l = 2$
5	1.3	
4		2.0
3	61.5	12.6
2	321.5	

TABLE V Labeling Schemes Used in XPS and X-Ray Spectroscopy

XPS	$1s$	$2s$	$2p_{1/2}$	$2p_{3/2}$	$3s$	$3p_{1/2}$	$3p_{3/2}$	$3d_{3/2}$	$3d_{5/2}$
X-Ray	K	L_I	L_{II}	L_{III}	M_I	M_{II}	M_{III}	M_{IV}	M_V

The intensity ratio of the two doublet lines is given by

$$\mathbf{I}_{l+1/2}/\mathbf{I}_{l-1/2} = (l+1)/l \tag{17}$$

The line with the higher intensity is always found at lower binding energy (see Fig. 14). Deviations from this ideal intensity ratio exist because of the effects discussed in Section II.E: one of the lines can lose more intensity to satellites than the other.

Spin–orbit coupling is less important for the valence electrons because l is no longer a good quantum number. Only for molecules that contain at least one three-fold or higher symmetry axis, the angular momentum around this axis can be described by a quantum number $l \geq 0$. If the angular momentum is nonzero, the angular motion can still couple to the spin. The intensity ratio in this case is 1:1 for the two doublet lines. In the valence region spin–orbit interaction is observed only when at least one atom with $Z > 10$ is present in the molecule. For light atoms, the doublet splitting due to spin–orbit coupling is too small to be resolved in normal UPE spectra. Spin–orbit interaction is also of great importance for the band structure of solids.

For states that contain more than a single unpaired electron (free radicals, atoms with partially filled d- or f-shells, $nh(n\text{-}1)p$ states, double hole states created by an Auger process), multiplet splitting can become extremely complicated. Splittings of this type are observed in the XPE spectra of Co and Ni ($2p$ signals), Mn and Cr ($3s$ signals), and the rare earths ($4s$ signals). Multiplet splittings can be used as a diagnostic tool as it distinguishes, e.g., the spectrum of nickel(II)oxide (Figure 10b) quite conclusively from nickel(II)hydroxide, where multiplet splitting does not occur. Auger signals usually consist of multiple lines, which are often only partially resolved. This can cause considerable difficulties in the proper determination of Auger parameters.

III. INSTRUMENTATION

A PE spectrum is measured with an instrument known as a photoelectron spectrometer. Figure 13 shows a sketch of

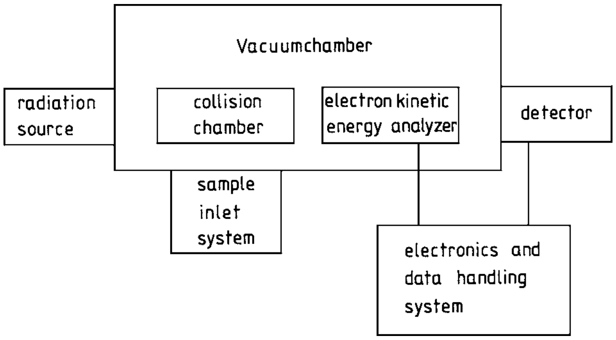


FIGURE 13 Scheme of a photoelectron spectrometer.

the basic structure of this instrument, which is the same whether the instrument is designed for gases or for solids. The technical details of the instrument, however, differ considerably depending on its application.

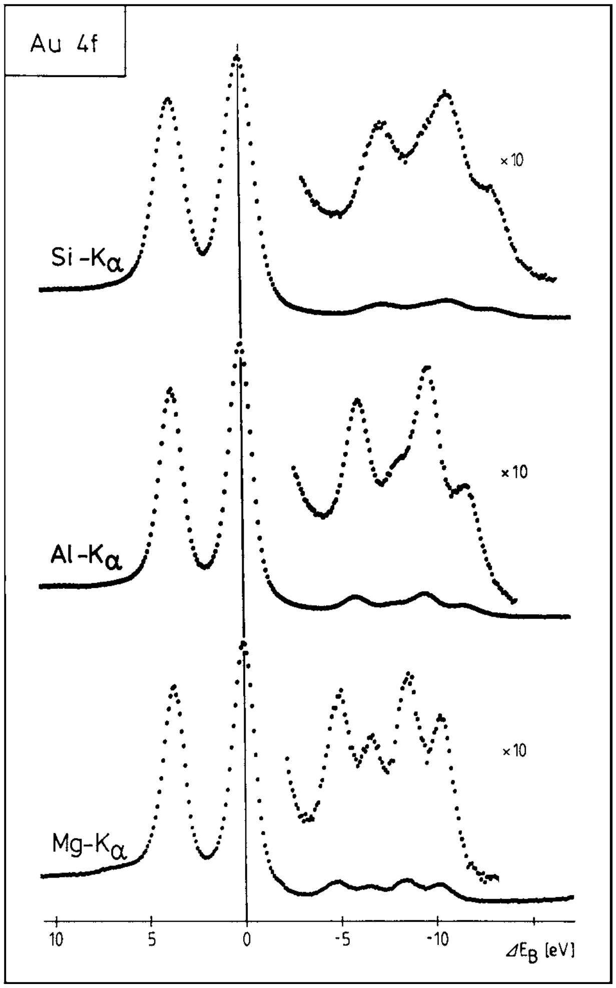


FIGURE 14 Au 4f doublet measured with different excitation sources.

A. Vacuum Requirements

Since PES measures electrons, the whole path between the region where photoionization takes place and the detector must be kept at a pressure of about 10^{-5} torr or less to prevent collisions with the rest gas. For an exciting radiation with $h\nu > 11.3$ eV, no material exists that is sufficiently transparent to serve as a window. Therefore, the path from the radiation source to the photoionization region (PIR) must also be kept under vacuum.

Photoelectron spectrometers are very sensitive to contamination of their inner surfaces. Deposition of the sample or its decomposition products on these surfaces can lead to local charges that strongly influence the path of the electrons. Therefore, the vacuum must not only provide a sufficient mean free path for the electrons but also prevent contamination as much as possible. Since even heavy pumping does not always prevent contamination, all vital parts of a PE spectrometer should be bakeable to remove contamination.

Photoelectron spectrometers designed for the study of solids require even better vacuum conditions. Because of the surface sensitivity of PES (see Section I.E and Fig. 4), any contamination of the sample surface contributes strongly to the measured PE spectrum. For a gas with a sticking coefficient of 1 (the sticking coefficient is the probability that a particle hitting the surface will remain on the surface), a clean surface is covered by a monolayer if 1 Langmuir of gas is offered to the surface (1 Langmuir is an exposure to 10^{-6} torr · sec). To maintain a reasonably clean surface for $^1/_2$ hr (about 10^3 sec), a vacuum better than 10^{-9} torr is necessary in the collision chamber. If well-defined adsorbates in the mono- or submonolayer range are the object of investigation, the vacuum in the collision chamber must be even better (down to 10^{-11} torr). Spectrometers designed for this type of investigation are therefore built in ultrahigh-vacuum (UHV) technology.

B. Radiation Source

The type of radiation source strongly depends on the kind of PE spectroscopy that is performed. With UPS the commonly used radiation source is a resonance lamp, usually operated with helium. The radiation emitted from this source results from the transitions $^1P(1snp) \rightarrow {}^1S(1s^2)$ (compare Table VI). The radiation is produced by a high-voltage, direct-current discharge in a capillary, a high-current arc discharge with a heated cathode, or a microwave discharge. The helium pressure is usually a few hundred millitorrs. Since this is higher than the pressure maintained in the collision chamber, differential pumping is necessary. The lamps have an intensity of about 10^{12} photons/cm · sec and a half-width of 10–15 meV,

TABLE VI Sources of Exciting Radiation

Source		Energy (eV)	Width (eV)
He	I	21.22	0.001
	I_α	21.2182	
	I_β	23.0848	
	I_λ	23.7423	
	II	40.81	0.01
	II_α	40.8140	
	II_β	48.3718	
	II_λ	51.0170	
	II_δ	52.2415	
Ne	I_α	16.6709	0.001
		16.8482	
	II_α	26.8141	0.01
		26.9110	
Ar	I_α	11.6237	
		11.8282	
Y	M_ζ	132.3	0.5
Zr	M_ζ	151.4	0.8
Ti	M_ζ	452.2	
Mg	K_α	1253.6	0.7
Al	K_α	1486.7	0.8
Si	K_α	1739.9	0.9
Zr	L_α	2042.4	1.7
Ag	L_α	2984.4	2.6
Ti	K_α	4510.9	2.0
Cr	K_α	5417.0	2.1
Cu	K_α	8047.8	2.5

corresponding to about 100 cm^{-1}. The half-width, which determines the resolution obtainable in the spectrum, can be reduced by altering the operating conditions, but at the cost of rapid intensity loss. The numbers given above are a compromise between intensity and half-width. Under normal operating conditions about 98% of the emission consists of the He I_α line at 504 Å. Thus, the helium resonance lamp is a source of fairly monochromatic radiation. If the lamp is operated at lower helium pressure and higher voltage, it also emits He II radiation, which is the radiation emitted from He^+. The strongest emitted line is the He II_α line at 40.8 eV (compare Table VI), and under optimal conditions up to 40% of the emitted radiation is He II_α. However, He I radiation is also present and has to be removed by optical techniques if pure He II_α radiation is needed. Under the conditions for optimal He II emission, the total intensity of the emitted radiation is lower than the intensity under the standard conditions for He I operation.

Resonance lines from other atoms can also be used for excitation. Data for some of these lines are included in Table VI. Since there is usually more than one intense line and since all the energies are lower than He I_α, these lines are used only for special investigations. However, some of these lines can show up in the helium discharge if the helium is not of very high purity. Because of the lower energy of the corresponding excited states, the relative intensity of the impurity lines is much higher than the concentration of the impurity itself.

For higher excitation energies, the only easily available monochromatic radiation comes from characteristic X-ray emissions. Because of the short lifetime of the involved core hole states (10^{-16} sec), the half-width of the emitted radiation is much larger than for the resonance lines discussed above and increases rapidly with the ordering number. The number of applicable photon sources is rather large and some of the more viable examples are listed in Table VI. The standard materials used in most instruments are aluminum and magnesium, often in a twin anode configuration, which provide K_α radiation at 1486.6 and 1253.6 eV, respectively. The filament is kept at ground while the anode is held at a high potential (12–20 kV). Twin anodes have two filaments which allow changes between Mg K_α and Al K_α by simply switching the power from one filament to the other.

Figure 14 shows the 4f doublet of polycrystalline gold measured with Si, Al, and Mg K_α radiation. The intensity obtained with characteristic X-ray lines is about two orders of magnitude lower than that of the resonance lines. High-power X-ray sources operating with a rotating anode have been designed for this reason. To gain spectra from deeper core levels, sources with higher energy are useful. However, the use of such sources is not without penalties in the form of decreasing spectral resolution and reduction in sensitivity for the very light elements.

Like the resonance lines, the X-ray lines are always accompanied by satellites with intensities of up to 10% of that of the main line. This must be considered in the evaluation of spectra. For example, the small structures seen at the low binding energy side of the photoelectron lines in Figs. 1a, 10c, and 14 are due to the satellites of the exciting radiation. In addition to the satellites, there is always a bremsstrahlung continuum underlying the characteristic X-ray lines. This continuum can be reduced by inserting a thin metal foil (the nature of which depends on the anode material) between the X-ray source and the collision chamber. The foil must be sufficiently thin and cannot be used to maintain a reasonable pressure difference between the source and the chamber.

To reduce the half-width of the characteristic X-ray lines, X-ray monochromators have been designed in which bent crystals are used to disperse the radiation. The method depends on the dispersion of X-rays by diffraction, as predicted by the Bragg equation

$$n\lambda = 2d \sin\varphi \qquad (18)$$

where n is the diffraction order, λ the X-ray wavelength, d the crystal spacing, and φ the Bragg angle. Anode, crystal, and sample surface are all arranged on the circumference of a so-called Rowland circle, usually 0.5 m in diameter. The size of the monochromator influences the size of the X-ray footprint, which is about 0.05% of the diameter of the Rowland circle. Commercial monochromators, which are now found in many XPS instruments, all rely on quartz as the 110 plane fulfills the required conditions for Al K_α radiation. With slight adjustments of anode and crystal position such a monochromator can also be used for Ag L_α ($n=2$) or Ti K_α ($n=3$). Linewidths down to 0.3 eV have been reached for Al K_α with this technique. The use of a monochromator has the additional advantage that the K_α satellites and the bremsstrahlung continuum are removed. The spectrum shown in Fig. 9 was obtained with a monochromator equipped instrument.

For the investigation of deeper lying core levels, which are not accessible with Al K_α radiation (1483.6 eV), like the $3d_{5/2}$ core level of Yb in YbP at a binding energy of 1527.8 eV, monochromatized Si K_α radiation is very useful. Monochromatization is again achieved with quartz (110) crystals. The Bragg angle for the Si K_α energy of 1740 eV is 56.85°; the diameter of the Rowland circle is 650 mm.

Synchrotron radiation has become increasingly important as exciting radiation. When charged particles travel along a bent path with a velocity near the speed of light, as occurs in an electron synchrotron or a storage ring, an intense beam of light is emitted tangential to the path. This light has a smooth continuous energy distribution extending far into the vacuum UV and is strongly polarized in the plane of the ring. Figure 15 shows the brilliance (number of photons emitted per square millimeter, second and steradiant) of the radiation provided by the new Berlin Synchrotron Radiation Facility BESSY II. Below the cut-off energy, which depends on the design of the synchrotron, the intensity is by two or three orders of magnitude higher than that of conventional line sources. Special devices, called *wigglers* and *undulators*, lead to a further intensity increase of two to four orders of magnitude (see Fig. 15). Such devices are now available at all synchrotron radiation facilities.

To use synchrotron radiation as an excitation source in PES it is necessary to select a small energy range from the continuous synchrotron radiation. This is done with different types of monochromators, which are provided by the synchrotron radiation facility. Great progress has been achieved over the last two or three decades in the construction of the necessary far UV and soft X-ray monochromators. The tunability of the excitation energy and the polarization of the synchrotron radiation add advantages that

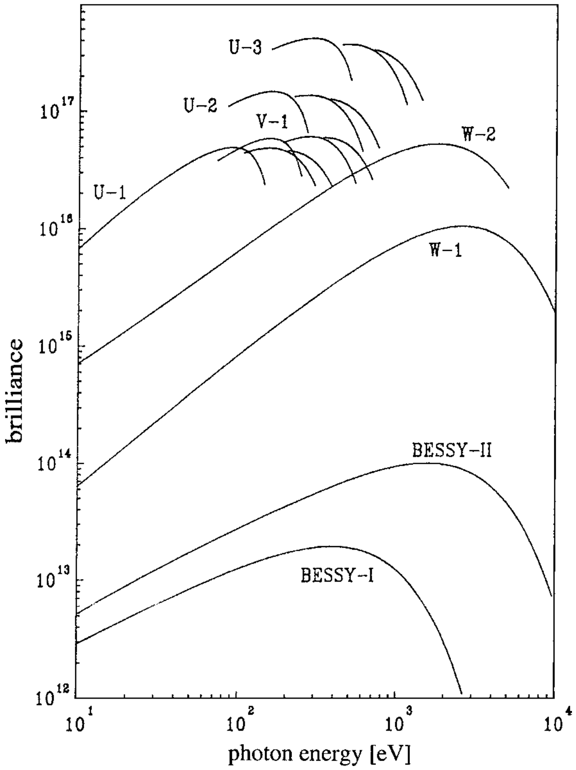


FIGURE 15 Brilliance $[in \frac{photons}{\sec(mm \cdot mrad)^2\, 0.1\%\, BW}]$ of the synchrotron radiation obtained from BESSY I (now closed) and BESSY II in Berlin. The curves for wigglers (W) and undulators (U, V) give an impression of the increase in brilliance that is gained with these devices. [From BESSY II (1986). "Eine optimierte Undulator/Wiggler-Speicherring Lichtquelle für den VUV- und XUV- Spektralbereich," p. 20, BESSY, Berlin.]

opened new dimensions for a great variety of experiments. One such example is shown in Section IV.B.

C. Collision Chamber and Sample Inlet System

For the measurement of gaseous samples the target gas is introduced through a capillary about 1 mm in diameter. The amount of gas entering the collision chamber is regulated by a leak valve. The beam of molecules coming from the capillary and the radiation coming from the source, which is usually a beam with a diameter of about 1 mm, can be either parallel or perpendicular to each other. The former arrangement leads to a rodlike photoionization region, the latter to a more pointlike one. The gas pressure in the PIR is about 10^{-3} torr. The photoelectrons created are usually observed perpendicular to the two incoming beams. This avoids having the beam of target gas point directly to the entrance of the kinetic energy analyzer and helps minimize contamination of the surfaces of the analyzer.

Vapors from liquids usually can he measured in the same way as gases. If the vapor pressure is lower than 10^{-3} torr, the liquid must be heated. To prevent condensation in the inlet system this should be heated, too.

To determine a gas-phase spectrum from a sample that is solid under normal conditions, the sample must be volatile. If the volatility is fairly high, the same inlet can be used as for liquids. For samples with low vapor pressure, a direct inlet system can be used. A small capillary is filled with the sample and placed in a heatable sample holder, and the opening of the capillary is then brought close to the PIR. Molecules evaporating from the capillary reach the PIR directly, and difficulties with deposition at narrow or cold parts of the inlet system do not arise. In addition, the molecules do not come into contact with heated metal parts, which often leads to catalytic decomposition. The amount of substance needed for the measurement of a gas-phase PE spectrum is about 20 mg, and it cannot be recovered.

To study photoemission from a solid, the sample must be brought directly to the ionization region. When the UHV requirements are not too high (up to 10^{-10} torr), the sample can be introduced through a lock. As extreme UHV conditions are required for most surface investigations, a fully metal sealed vacuum system must be used. The sample is usually mounted on a manipulator that can be operated from the outside and is inserted through a system of two or more separate vacuum chambers. This prevents the pressure in the analysis chamber from rising upon introduction of a sample and thus guarantees a high operational capacity.

If the sample is a conductor, it is brought in direct electrical contact with the sample holder. Metallic indium is often used to provide the necessary electrical and thermal contact between sample and sample holder. This is especially convenient because samples that are small particles can be pressed directly into the indium. If the sample is an insulator, even pressing it into indium does not always prevent charging. In this case, charge compensation can be achieved by using a floodgun to bombard the sample with low energy electrons. Powder samples can be mounted on double-stick tape or pressed into a tablet, similar to the preparation used in infrared spectroscopy. Soluble samples can be dissolved and a drop of the solution brought to the surface of a metallic sample holder and evaporated. The latter method can yield a thin film, which is less sensitive to charging.

When a sample is brought into the vacuum, its surface is always contaminated, usually with O_2, H_2O, CO_2, and hydrocarbons. In many cases, the sample surface itself has a different chemical constitution than the bulk. The surface of most metals, for example, consists of oxides. In nearly all cases, the sample surface must therefore be cleaned under UHV conditions. Among the techniques most widely used for the preparation of suitable surfaces are

1. Bombardment of the surface with rare gas ions with a kinetic energy of a few hundred to a few thousand electron volts
2. Heating of the sample, sometimes in the presence of hydrogen gas (to reduce oxides)
3. Mechanical preparation of fresh surfaces within the UHV.

Which preparation technique is most suitable always depends on the nature of the sample. For example, ion bombardment can change the surface constitution and heating can lead to phase transitions. To prevent contamination of the collision chamber, the cleaning and preparation procedure is usually carried out in a separately pumped chamber. The sample is then transferred from the preparation chamber to the collision chamber.

D. Electron Energy Analyzer

The electron energy analyzer is the heart of a PE spectrometer. It is here that the electrons are discriminated with respect to their kinetic energy. The most important features of an electron kinetic energy analyzer (EKEA) are sensitivity and resolution. High resolution and high sensitivity contradict each other, and a given analyzer will always be a compromise between them. As chemical shifts are often small (<0.5 eV) the spectral resolution, ΔE, of an analyzer used in XPS should not exceed a few tenths of an eV. ΔE is mainly determined by three contributions: the analyzer resolution ΔE_A, the linewidth of the exciting radiation ΔE_X, and the intrinsic width of the photoelectron line ΔE_P. If we assume Gaussian distributions for all three, ΔE can be expressed as

$$\Delta E = \left(\Delta E_A^2 + \Delta E_X^2 + \Delta E_P^2\right)^{0.5} \tag{19}$$

As discussed in Section III.B, ΔE_X is at best 0.3 eV in XPS, when using a monochromated X-ray source, and 10 meV in UPS. To minimize the influence of the analyzer, ΔE_A should by no means exceed these values. In both cases, this corresponds to a relative resolution of at least 1:1000.

Modern EKEAs usually consist of a complex combination of electron transfer lenses and a hemispherical electrostatic deflection analyzer. Figure 16 shows a sketch of the analyzer and input lense system presently used on a VG ESCALAB® 220iXL produced by VG Scientific®. The input lens captures the maximum possible number of emitted phototelectrons and transfers them to the deflection analyzer.

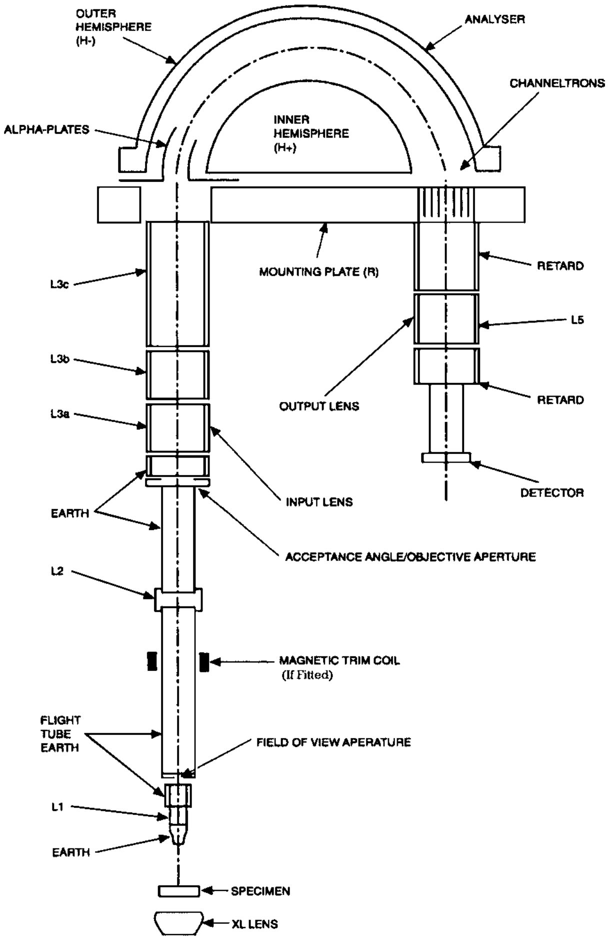


FIGURE 16 Sketch of the analyzer/detector system of the VG ESCALAB 220iXL® produced by VG Scientific®.

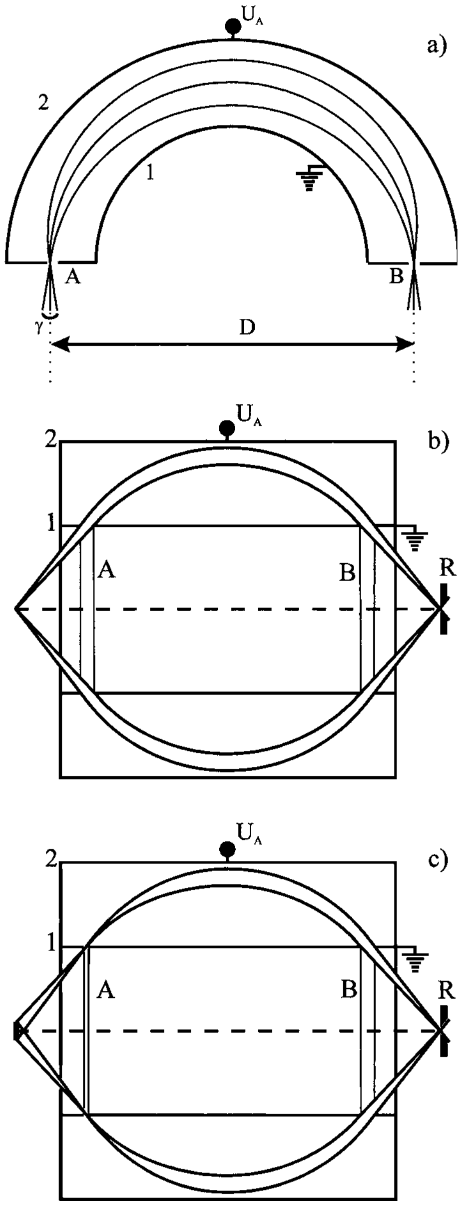


FIGURE 17 Schematic cross section of (a) a hemispherical analyzer, (b) a cylindrical mirror analyzer with axis–axis-type focusing, and (c) ring–axis-type focusing. 1 and 2—inner and outer electrodes; A and B—aperture windows; R—receipt diaphragm.

Figure 17a shows the principle of a hemispherical analyzer. At a given voltage U_A, electrons that enter through the aperture A with a certain kinetic energy (the pass energy) can leave through the exit aperture B. The relation between pass energy and U_A is determined by the distance D. If electrons that do not enter perpendicular to the entrance aperture leave through the exit aperture, the analyzer is *focusing*. If focusing conditions are met only within one plane the analyzer is called single focusing. If the conditions are met in two dimensions the analyzer is double focusing. The hemispherical analyzer is first-order double focusing, thus it accepts electrons from a solid angle γ. The larger the γ, the higher the sensitivity at a given pass energy.

The resolution of the analyzer depends on its geometric parameters, that is, the radius of the hemispheres, the area of entrance and exit aperture, the angle γ, and the pass energy. For a fixed γ, the resolution increases with decreasing pass energy until a limit (usually a few eV) where stray fields become important. For this reason electrons are retarded before they enter the deflection analyzer. This is usually done in the last part of the lens system. However, decreasing the pass energy also decreases the sensitivity. Therefore modern instruments allow choosing of different pass energies and thereby the selection of either the highest possible sensitivity or the highest possible resolution or an appropriate compromise between the two. One consequence of Eq. (19) is that reducing the pass energy

will make the terms E_A and E_P more dominant. For an intrinsically wide peak, reducing the pass energy past a certain point will not reduce the achieved width further and will only serve in reducing the sensitivity.

The combination of retardation and deflection allows two different modes of operation. In the constant analyzer energy (CAE) mode the pass energy is kept constant and the spectrum is scanned by a variation of the retarding field. In this mode the resolution ΔE_A is kept constant over the whole spectrum. In the constant retard ratio mode (CRR) the voltage difference between the hemispheres is proportional to the retarding field. In this mode the relative resolution $\Delta E/E$ is kept constant. The CAE mode is normally used for PES, the CRR mode for Auger spectroscopy.

An electrostatic deflection analyzer widely used in Auger spectroscopy is the cylindrical mirror analyzer (CMA). There are two different types of CMA, which correspond to two different regimes of the angular second-order focusing: the axis–axis and the ring–axis type focusing. A CMA with axis–axis type angular focusing, as sketched in Fig. 17b, is composed of two coaxial cylindrical electrodes. The inner electrode has two ring-shaped aperture windows, covered by a metallic grid. At the CMA exit a receipt diaphragm is placed, behind which a detector for charged particles is installed. A charged particle beam, after its deflection in the field is focused to the central opening of the receipt diaphragm and reaches the detector. The spectrum is recorded by varying the potential between the two cylindrical electrodes.

The major difference of a CMA with ring–axis type angular focusing (Fig. 17c) is a narrow ring-shaped slot in the inner cylindrical electrode that functions as an electron optical source. The CMA thus analyzes a beam of charged particles passing from the emitter surface to the ring-shaped opening.

The advantage of a CMA is its relatively high sensitivity due to its large γ; however, this is at the expense of resolution. Fitting a hemispherical analyzer with a high transmission electron transfer lens makes the sensitivity of this type of analyzer comparable to that of a CMA, with the added advantage of high resolution.

E. Detectors

Modern spectrometers generally use multichanneltron arrays similar to the one shown schematically in Fig. 18. The array is mounted in the output plane of the analyzer. The width of the array matches the gap between the two analyzer hemispheres. The exit aperture is defined by the area of the individual channeltron. The channeltrons act as highly sensitive amplifiers whose gain depends on the voltage applied across the detector. When a photoelectron

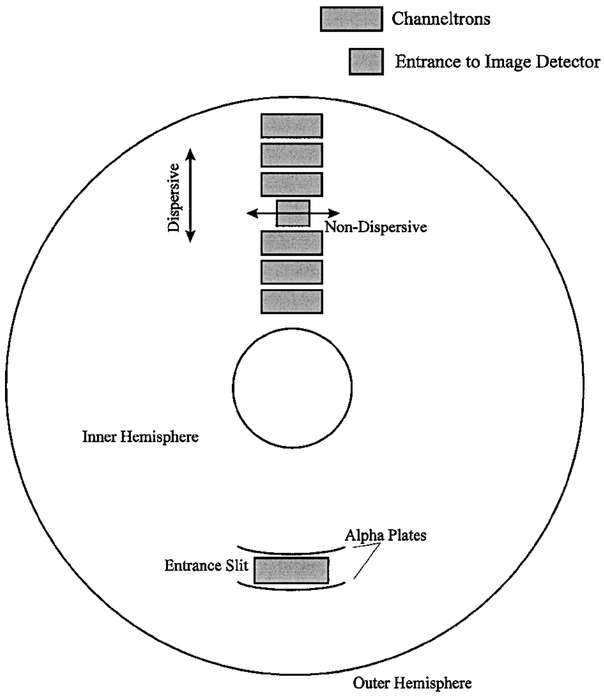


FIGURE 18 Rear view of the channeltron assembly used in the VG ESCALAB 220iXL®.

hits the resistive coating at the input of a channeltron, secondary electrons are produced. These secondary electrons are then accelerated to hit a deeper channeltron wall and produce another avalanche of electrons. This process finally produces a pulse of up to 10^8 electrons for every original incident electron. These pulsed output signals are fed to preamplifiers mounted just outside the vacuum and then onto the data system. Each channeltron detects the presence of electrons in a specific energy window. Thus the array simultaneously covers a range of different energies. These signals can then be combined in the software to produce a spectral trace.

Measurement of a PE spectrum with count rates down to a few electrons per second is time consuming. Statistical noise is always proportional to the square root of the number of electrons counted per unit time. If only 10 electrons are counted for a given kinetic energy, the uncertainty is about 30%. If 90% of the electrons are due to background (see Fig. 1c), the signal is lost in the noise. To detect the signal, the noise must be reduced to at least 5%, corresponding to a collection of about 400 electrons. A count rate of 10 per second leads to an observation time of 40 sec at this single kinetic energy. The time, which is necessary for the measurement of a certain energy range, is considerably reduced by the multiplex advantage of the channeltron array and the use of a high transmission electron transfer lens but can be longer than a few minutes when

analyzing weak features in the spectrum. Most modern instruments use a sweep technique in which the investigated energy range is measured within a few seconds, and the measurement is repeated at intervals. The data from each sweep are accumulated and stored in a data handling system. Well-designed sweep programs also take care of sensitivity differences between individual channeltrons in the array. The development of the spectrum can be watched on a screen, and the measurement can be terminated when the signal-to-noise ratio has reached an acceptable limit.

For spin-resolved PES (Section I.H), the detector must be sensitive to the direction of the electron spin. Such a detector was constructed in 1943 by Shull *et al.* based on the theoretical prediction by Mott, that the scattering of an electron from a nucleus with a high atomic number Z depends on the direction of the incident electron's spin, if the electron has an energy greater than 50 keV and is scattered through an angle greater than 90°. This Mott detector has been widely used to measure the spin polarization of electrons. In the 1980s, Mini Mott detectors were realized by the use of cylindrical or sperical retarding optics. Among them, the SPLEED and the low-energy diffusive scattering detector are of practical importance. The latter is preferable because the SPLEED demands frequent cleaning of the target crystal under UHV. The low-energy diffuse scattering detector is based on the spin–orbit interaction between an incident electron and a polycrystalline gold film that is used as scattering target, because gold is not chemically reactive and exhibits strong spin-orbit interaction due to its high atomic number Z. According to the **Ls** term of the spin–orbit interaction, the spin analyzer is sensitive to the transverse component of the spin polarization which is normal to the scattering plane, while the longitudinal component along the incident electron direction cannot be observed.

F. Magnetic Shielding

Low-energy electrons are very sensitive to the stray magnetic fields that exist in any laboratory, such as the magnetic field of the earth. These stray fields strongly influence the trajectories of the traveling electrons. As a result, all parts of a PE spectrometer between ionization region and detector must be carefully shielded. This can be done with Helmholtz coils, which compensate the external fields, or with magnetically shielding materials like mu-metal. The latter type of shielding is used in most modern instruments, because shielding with Helmholtz coils is difficult for the dynamic stray fields produced by many types of laboratory equipment. Improper shielding leads to reduced sensitivity and resolution, as well as to asymmetric skewed lines.

G. Instrumentation for Angle-Resolved Photoelectron Spectroscopy

Instruments designed for angle-resolved measurements must meet two specific requirements: (1) The analyzer must be movable around the ionization region and (2) the acceptance angle at the first input lens of the analyzer system must be small enough (usually a few degrees) to achieve angular resolution. The requirements for an angle-resolving spectrometer differ depending on whether the sample is a gas or a solid.

The angular distribution for atoms or molecules in the gas phase is fully described by the asymmetry parameter β (Section I.F). To measure β, the angle θ between the incoming radiation and the outgoing electrons must be varied [Eq. (6)]. It is sufficient for this type of measurement to move the analyzer in one plane.

For solid samples there are three degrees of freedom: the angles Ω, θ, and ϕ defined in Section I.G. If the incident radiation is polarized, an additional degree of freedom is obtained because of the orientation of the polarization vector. Angle-resolved measurements with unpolarized radiation can be carried out with an instrument in which the analyzer can be moved in one plane, provided that the sample is mounted on a manipulator that allows adjustment of the direction of the surface normal. With this setting Ω, θ, and ϕ can be varied over a fairly wide range. However, a better range for all three angles is achieved if the analyzer can be moved in two planes that are perpendicular to each other. Because of the additional degree of freedom obtained with polarized radiation, it is necessary to use an instrument that allows movement of the analyzer in two dimensions in order to obtain all the available information.

Some limited information on the angular distribution of the electrons can be derived even with an instrument in which the angle between excitation source and analyzer is fixed: if the sample is mounted on a suitable manipulator, it is possible to vary the orientation of the analyzer to the surface normal while keeping the angle of incident radiation constant.

IV. SOME EXAMPLES

In this section we will discuss a few examples of the application of PES in different fields of research. The number of examples is too limited to allow more than a glimpse onto the broad and highly active field of PES. Further examples and other types of application can be found in the references cited in the bibliography. Current work in PES is published in a variety of scientific journals. A journal specifically devoted to this field is the *Journal of Electron Spectroscopy and Related Phenomena*, published by Elsevier, Amsterdam.

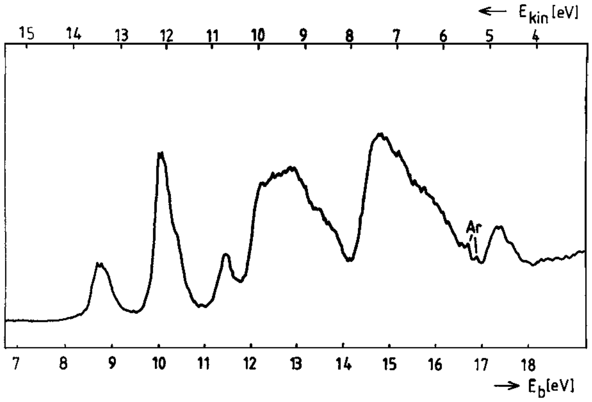


FIGURE 19 UPE spectrum of *trans*-stilbene.

A. Free Molecules

The type of investigation that can be performed with free molecules has been illustrated throughout this article by the example of formaldehyde: experimental binding energies can be compared to calculated orbital energies. We can try to understand differences in relaxation contributions, and, by inspection of the shape of the PE bands, we derive information on the bonding characteristics of different orbitals. This type of analysis is, however, limited to smaller molecules. When the molecules become larger, the number of orbitals in the valence region increases and the different bands overlap so strongly that a separate assignment is often impossible. An example of such a case is shown in Fig. 19. The bands between 11.8 and 18 eV are caused mainly by photoemission of electrons from CH bonds. Because of the large number of CH bonds in stilbene and the vibrational broadening of the single bands, this energy region of the spectrum can no longer be analyzed in detail. In the low-energy region, however, three well-separated bands are observed with an intensity ratio of approximately 1:3:1. These bands can be correlated to the five highest occupied orbitals of *trans*-stilbene, which are all π orbitals. This situation, where it is possible to assign several bands at the beginning of the PE spectrum of a larger molecule, is usually met in cases where the uppermost occupied orbitals are energetically well separated from the main body of the valence orbitals. Among these special orbitals are *n* orbitals, π orbitals, and the *d* orbitals of transition metal complexes.

Very often it is not the electronic structure of a single compound that is studied by PES. In most cases, it is the change in electronic structure caused by some variations in chemical structure that one wants to understand. This type of investigation is illustrated by the following two examples.

When stilbene is substituted in the *para* position with an electron-donating substituent such as a methoxy group, the π orbitals are expected to rise in energy, which in turn should lead to smaller binding energies. The influence of a substituent on a given orbital is proportional to the contribution of the atom on which the substituent is located to this orbital. The contributions of the different carbon atoms to the five highest occupied π orbitals of *trans*-stilbene are indicated in Fig. 20. Orbitals π_5 and π_6 have negligible contributions at the *para* atoms and are expected to be relatively insensitive to *para* substitution. Comparison of the observed spectra (Fig. 21) to the expected band shifts leads to an unambiguous assignment, which cannot be derived from one of the individual spectra. For example, it is not directly obvious that the shoulder observed in the *trans*-stilbene spectrum at 9.5 eV corresponds to an ionization from π_4. However, if the series stilbene, *p*-methoxystilbene and p,p'-dimethoxystilbene is considered, it is clear that the orbitals hat have

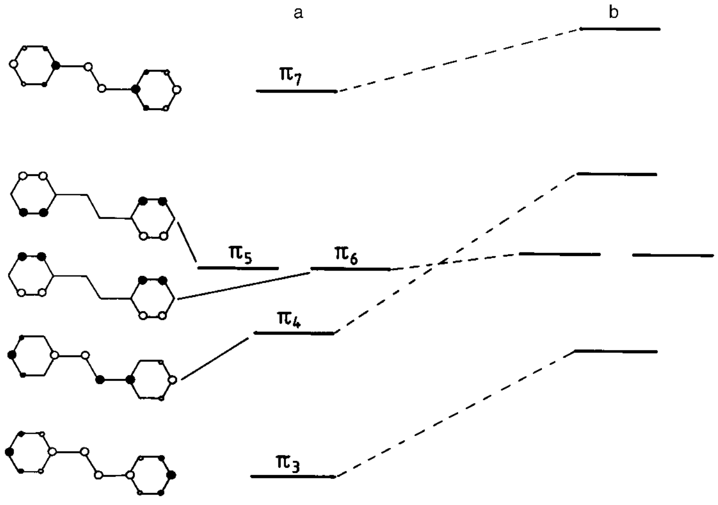


FIGURE 20 Orbital diagram for stilbene: (a) ordering as in *trans*-stilbene; (b) ordering as in *trans*-p,p'-dimethoxystilbene.

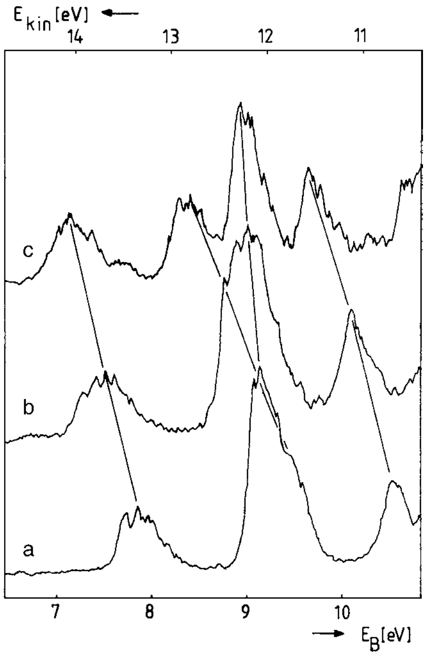


FIGURE 21 UPE spectra for (a) *trans*-stilbene, (b) *trans-p*-methoxystilbene, and (c) *trans-p,p*′-dimethoxystilbene.

contributions at the *para* atoms (π_3, π_4 and π_7) are shifted upward by nearly equal amounts. In contrast, the band corresponding to the nearly degenerate orbitals π_5 and π_6 moves very little.

The spectra of our second example are shown in Fig. 22. Cyclohexene shows a well-separated band at the onset of the PE spectrum with a vertical ionization potential of 9.1 eV. This band, which is not seen in cyclohexane, is undoubtedly related to the π orbital of the double bond. In contrast to the bands at higher energies, the π ionization band exhibits vibrational fine structure due to a strong coupling of the ionization process to the C-C stretching vibration of the double bond. In 1,4-cyclohexadiene we find two bands at the beginning of the spectrum (VIPs 8.8 and 9.8 eV) that must be attributed to ionizations from the π system. At first thought, one expects only one band in 1,4-cyclohexadiene, since the two double bonds are not conjugated. The splitting tells us that there is some interaction between the two nonconjugated π bonds. This interaction can be caused either by direct overlap, resulting from the nonplanar conformation of the system, or by an interaction with σ bonds of appropriate symmetry. The first type of interaction is called "through space" and the second type "through bond" interaction. In order to distinguish between these two types of interactions a careful theoretical analysis is necessary.

B. Energy Band Mapping

The quadratic relation between energy and momentum that holds for a free electron is no longer valid in a crystalline solid, where the electron moves under the influence of the periodic lattice potential. The relation between electron energy and electron momentum along certain directions within the Brillouin zone is called the "band structure" of the solid. Angle-resolved PES is presently the only method that allows an experimental determination of the band dispersion (the relation between energy and momentum) of occupied bands that do not lie close to the Fermi energy. This capability is due to the fact that the photon adds only negligibly to the electron momentum. Because of momentum conservation, the momentum of the created photoelectron (p_{PE}) is the same as that of the electron in the initial state band (p_i). The simultaneous fulfillment of energy conservation

$$E_{kin} = h\nu - E_B \tag{20}$$

and momentum conservation

$$\mathbf{p}_{\text{PE}} = \mathbf{p}_i \tag{21}$$

has consequences that can be understood from the simplified band structure shown in Fig. 23. As usual for band structures, the wave vector $\mathbf{k} = \mathbf{p}/\pi$ instead of the momentum is drawn at the abscissa. To a first approximation, the emitted photoelectron can be treated as a free electron, which yields

$$E_{kin} = \mathbf{p}_{\text{PE}}^2/2m = \mathbf{p}_i^2/2m = \pi^2\mathbf{k}^2/2m \tag{22}$$

Thus, selecting a specific kinetic energy corresponds to selecting a specific k value. All processes that lead to the proper momentum must lie on a vertical line defined by this k value. Conservation of energy [Eq. (20)] allows photoemission to occur only in cases where the binding energy with respect to the vacuum level and the kinetic energy of the photoelectron add up to the excitation energy $h\nu$ (see Fig. 23). For a given $h\nu$, photoemission is possible only for selected binding energies corresponding to a point within a given band. When the excitation energy is changed, the simultaneous fulfillment of energy and momentum conservation selects other points within the band and peaks in the energy distribution curve will occur at other binding energies. An example of such a measurement is shown in Fig. 24, where photoemission from a Pt(111) surface is studied in the valence region with excitation energies between 12.7 and 23.9 eV.

The analysis outlined above is complicated by the fact that for kinetic energies up to about 50 eV the created photoelectron is not really a free electron. It is excited into a final state whose dispersion may deviate considerably from the parabolic form of Eq. (22). A second type of complication results from the fact that, to be measured,

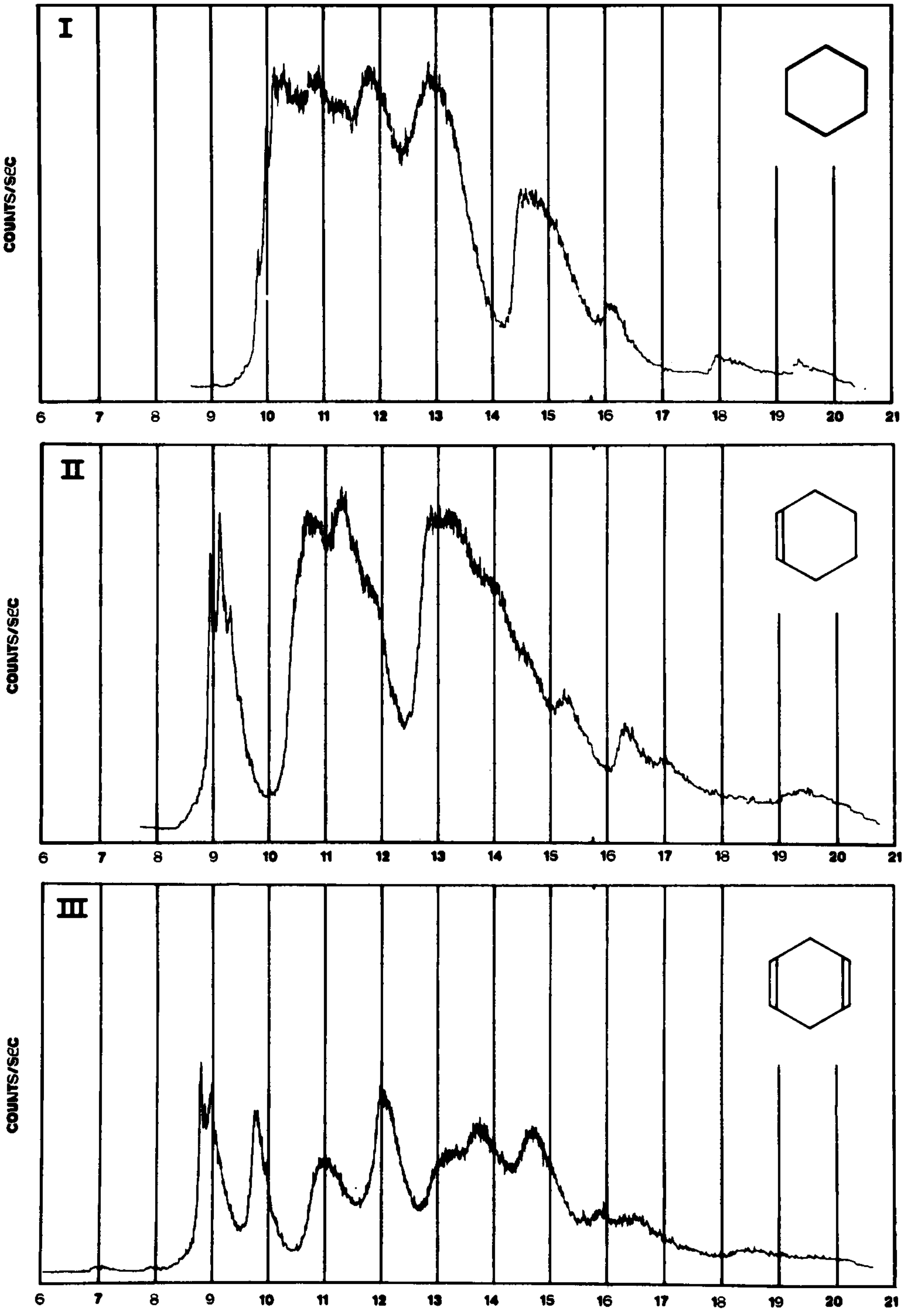


FIGURE 22 UPE spectra of cyclohexane, cyclohexene, and 1,4-cyclohexadiene. [From Bischof, P., *et al.* (1969). *Helv. Chim. Acta* **52,** 1745.]

the electron must escape into the vacuum, and in going from the interior of the crystal through the surface into the vacuum only the parallel component $k_{\parallel}$ of the momentum is conserved and not the component perpendicular to the surface (see inset in Fig. 25). A variation of the polar angle θ selects different parallel components

$$k_{\parallel} = \left(2mE_{kin}/\pi^2\right)^{-1/2} \sin\theta \qquad (23)$$

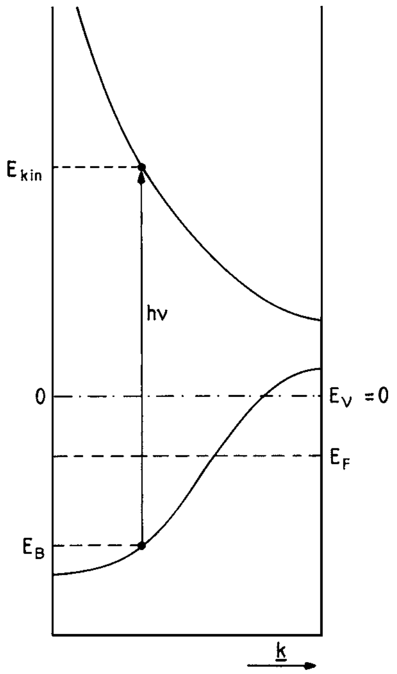


FIGURE 23 Scheme of a band structure showing the constraints imposed by simultaneous fulfillment of energy and momentum conservation.

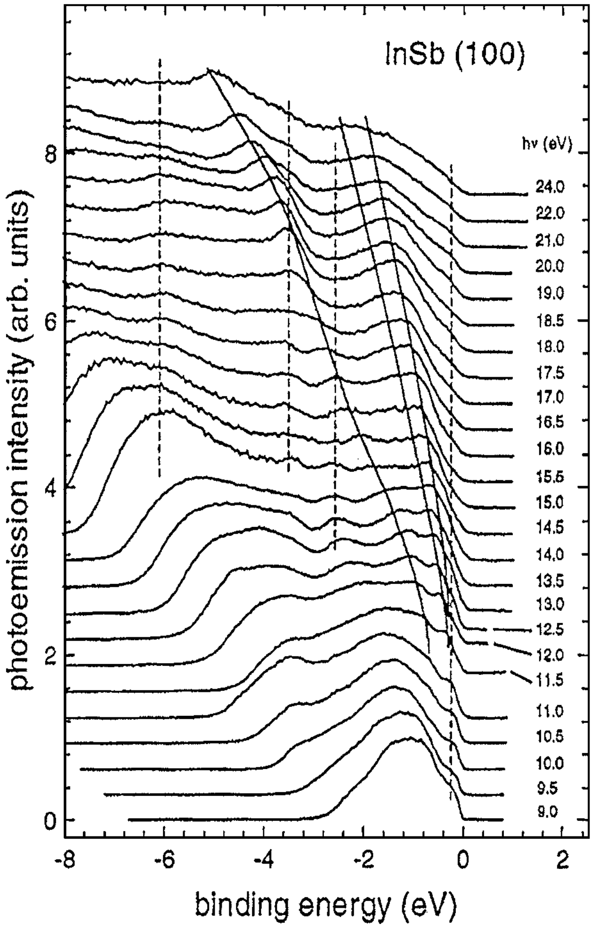


FIGURE 24 Energy distribution curves obtained with different excitation energies from a InSb (100) surface under normal emission ($\theta = 0°$). The binding energy is given with respect to the Fermi energy. The photon energy is indicated on the right side. Lines are drawn to guide the eye only. [From Jung, Ch., and Bressler, P. R. (1996). *Electron Spectrosc.* **78,** 503.]

Consequently, energy distribution curves measured under different polar angles, but using the same excitation energy $h\nu$, strongly depend on θ (Fig. 25). For two-dimensional lattices which occur in well-ordered adsorbates or in layer compounds, the band dispersion is only a function of $k_{\|}$, and the unknown change in the vertical component $k_{\perp}$ is of little importance in the analysis of the energy distribution curves. For three-dimensional lattices a variety of methods has been developed to overcome the problems resulting from the change in $k_{\perp}$. An experimentally determined band structure of copper is shown in Fig. 26. The dashed lines show the results of theoretical predictions. To derive the experimental band structure the polar angle θ and the excitation energy $h\nu$ had to be varied. Because of the latter, this kind of investigation depends on the use of synchrotron radiation.

C. Atomic Charges

In Section II.A we showed how core orbital energies can be expressed in the point charge approximation (Eq. (9)]. By combining Eq. (9) with Eq. (8) we obtain

$$\Delta E_B(i) = k(A, i)\Delta q_A + \Delta V(q_B) \qquad (24)$$

which connects core electron binding energy shifts ΔE_B with variations in atomic charges. The most drastic approximation is to correlate binding energy shifts only with the charge at the atom that holds the orbital. This simple approximation was widely used in the early days of XPS. However, it is an oversimplification because it neglects not only the off-atom potential $V(q_B)$ but also final state effects. A much better result is obtained when we use the method outlined at the end of Section II.B. From the proper combination of Auger kinetic energies and binding energies we can derive a close approximation to the relaxation contribution ΔR. Combination of Eqs. (14) and (12) allows us to take final state effects into account. We then derive quasi-experimental $\Delta\varepsilon_i$ values that can be used in connection with Eq. (9) and theoretically calculated atomic charges. An example of such a study is shown in Fig. 27. The numbers in the figure refer to the phosphorus compounds listed in Table VII, which shows the experimental data (Auger kinetic energy shift and binding energy shifts relative to PH_3), the final state

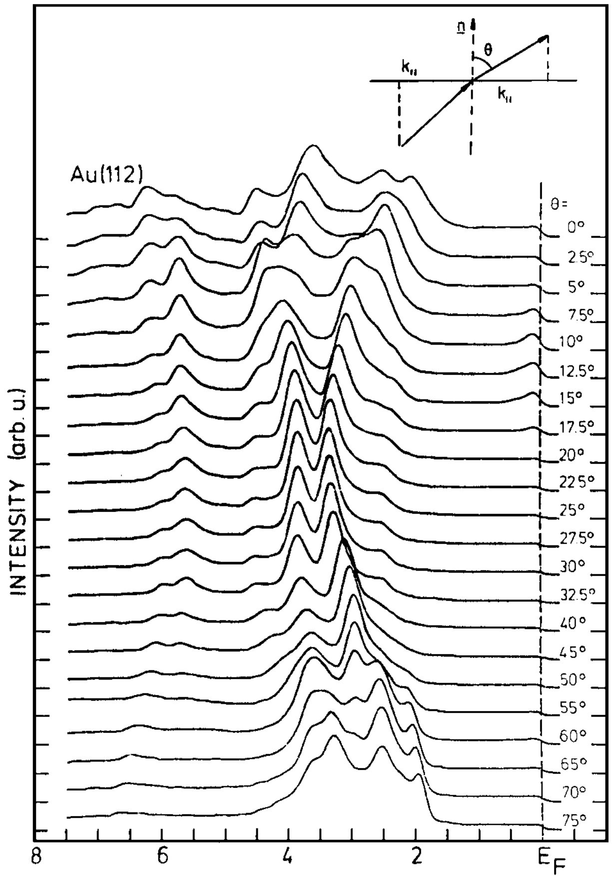


FIGURE 25 Energy distribution curves measured under different polar angles θ for an Au(112) surface. The binding energy is given with respect to the Fermi energy. The excitation energy is 16.85 eV. [From Heimann, P., Miusga, H., and Neddermeyer, H. (1979). *Solid State Commun.* **29,** 463.]

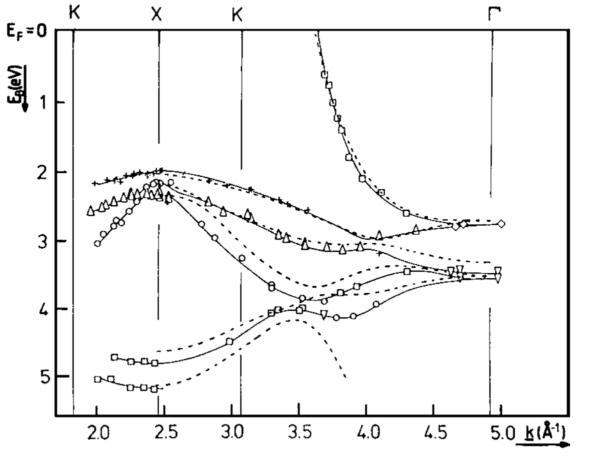


FIGURE 26 Experimentally determined band structure of copper along the Σ direction. [From Baalmann, A., *et al.* (1985). *Solid State Commun.* **54,** 583]. The dashed lines represent theoretical results derived by Eckardt, H., Fritsche, L., and Noffke, J. (1983). *J. Phys. F* **14,** 97.

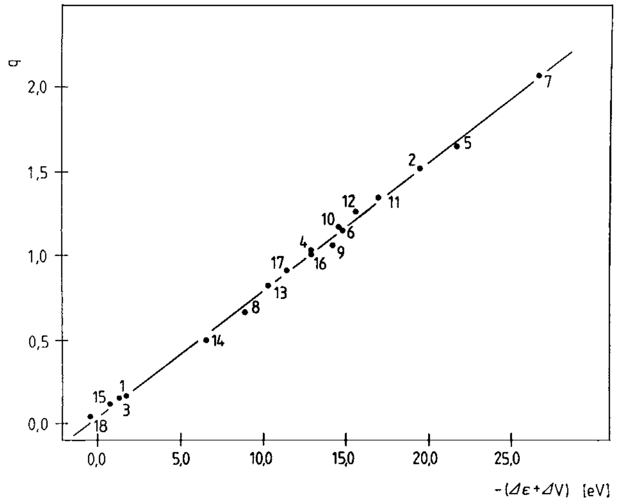


FIGURE 27 Correlation between calculated atomic charges and experimentally derived initial state contributions (see text for details). The numbers refer to the compounds listed in Table VII.

relaxation contribution ΔR, and the "experimentally" derived orbital energy shift $\Delta\varepsilon$ for the $2p$ orbital of phosphorus. Nearly all the ΔRs are positive, indicating that in almost all investigated compounds the final state relaxation is larger than in PH_3. Only PF_3 and OPF_3 exhibit small

TABLE VII Evaluation of Initial ($\Delta\varepsilon$) and Final State (ΔR) Effects from Experimental Binding Energy (P 1s and P 2p) and Auger Kinetic Energy ($KL_{2,3}L_{2,3}$) Shifts[a,b]

Compound	No.	(1*s*)	(2*p*)	E_{kin}^{Au}	ΔR	$\Delta\varepsilon$
PH_3	1	0.00	0.00	0.00	0.00	0.00
SPF_3	2	6.20	5.15	−2.98	0.56	−5.71
PMe_3	3	−0.98	−1.10	3.30	1.04	0.06
$SPCl_3$	4	4.21	3.57	0.47	1.70	−5.27
OPF_3	5	6.96	5.91	−5.14	−0.14	−5.77
$OPCl_3$	6	4.70	4.01	−0.72	1.30	−5.31
PF_5	7	8.55	7.33	−5.93	0.09	−7.42
PCl_3	8	3.28	2.79	0.39	1.35	−4.14
PF_3	9	5.48	4.72	−4.62	−0.33	−4.39
$(CH_3O)_2PSCl$	10	3.60	2.78	0.95	1.46	−4.24
$(CH_3O)_3PO$	11	3.34	2.54	0.11	0.92	−3.46
$(CH_3O)_3PS$	12	3.26	2.39	0.95	1.24	−3.63
$(CH_3O)_3P$	13	1.82	1.19	1.03	0.80	−1.98
CH_3PCl_2	14	1.89	1.53	1.33	1.25	−2.78
PEt_3	15	−1.48	−1.53	4.48	1.45	0.08
$(CH_2Cl)POCl_2$	16	3.77	3.18	0.10	1.34	−4.52
CH_3PSCl_2	17	3.19	2.59	1.44	1.71	−4.30
$P(CF_3)_3$	18	1.74	1.70	1.16	1.41	−3.11

[a] Experimental data from Sodhi, R. N., and Cavell, R. G. (1983). *J. Electron Spectrosc.* **32,** 283.

[b] All energies are in electron volts.

negative relaxation shifts. As suggested in Section II.B, the final state relaxation is largest for the molecules that contain easily polarizable substituents. In Fig. 27, the $\Delta\varepsilon$ values obtained from the experimental binding energy shifts and the experimentally estimated relaxation contribution are compared to the charge at the phosphorus atom and the off-atom potential, both calculated by a semiempirical quantum chemical method. The excellent correlation obtained for this example tells us that the theoretical method used is suitable for predicting reasonable atomic charges.

The energies shown in Table VII were obtained from gaseous samples. For solids, an additional problem arises that is connected with the reference problem (Section I.E). For solids, the binding energies are usually referred to the Fermi level and not to the vacuum level. The binding energy shifts which appear in Eq. (12), however, refer to the vacuum level. Binding energy shifts referred to the Fermi level and those referred to the vacuum level are equal only when the work function is the same for both samples. If this is not the case, the work function difference, which in principle is also an experimentally obtainable quantity, must be taken into consideration.

D. Investigation of Adsorbates

When a molecule is adsorbed at the surface of a solid it can be either physisorbed or chemisorbed. We speak of "physisorption" when the bonding is mainly caused by van der Waals interactions and of "chemisorption" when we have some type of a chemical bond between adsorbate and substrate. However, there is a more or less continuous transition between both types of bonding. The nature of the bonding between adsorbate and substrate is very important for the understanding of reactions that take place in heterogeneous catalysis. Catalytic processes of this type have great technical importance but often they are not fully understood on a molecular level. Because of its extreme surface sensitivity, PES is a powerful tool in achieving such an understanding. The applications of PES to the study of adsorbates are too widespread to be reviewed here. We only show two examples to give an impression of how these studies can be carried out.

For a variety of adsorbate systems such as the adsorption of small molecules (CO, N_2, H_2O, etc.) on transition metal surfaces, a "molecular view" of the bonding often provides a useful description. In a molecular view we do not look at the substrate as a metal with band structure and all the typical solid-state properties. Instead, we look primarily at a few metal atoms to which the adsorbed molecule is bound. Such a view immediately connects the bonding in an adsorbate to the bonding in a metal complex. Thus, studying adsorbates together with related complexes can

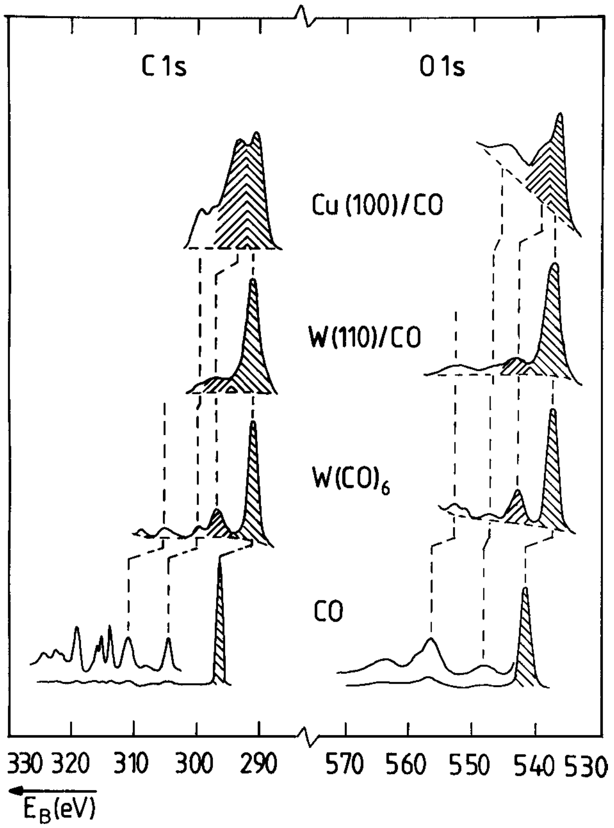


FIGURE 28 Comparison of C 1*s* and O 1*s* spectra for free CO, $W(CO)_6$, and CO adsorbed on W(110) and Cu(100). [From Freund, H. J., and Plummer, E. W. (1981). *Phys. Rev. B* **23,** 4859.]

contribute strongly to the understanding of the chemisorption bond. Figure 28 shows an example. The C 1*s* and O 1*s* spectra of gaseous CO are shown at the bottom. Weak satellite structures at the high binding energy side of the main lines are blown up in the figure in order to reveal their structure. The indicated satellites are connected with final ion states in which, in addition to the removal of a core electron, a second electron is excited from the π to the π^* orbital. When CO is adsorbed on the 110 surface of tungsten, a new satellite is observed closer to the main line. This new satellite increases strongly in intensity with decreasing strength of the chemisorption bond. For the system CO on copper (100) (at the top in Fig. 28), where CO is only weakly bound, this new satellite becomes very strong. What is the nature of this satellite and what does it tell us about bonding? The fact that the same satellite as observed for CO on W(110) is also observed in the metal complex $W(CO)_6$ reveals that this satellite cannot result from the special properties of metallic tungsten. It must be related to the local bonding between CO and a tungsten atom. The satellite is connected with a final state where, in addition to the core ionization, an electron is transferred

from an occupied d orbital of the tungsten atom to the π^* orbital of CO (charge transfer excitation). A detailed analysis reveals that the intensity of the satellite is a measure of the "backbonding" caused by the overlap of occupied d orbitals of the metal atom with empty π^* orbitals of the ligand. The similarities and differences observed in the satellite structure of the C 1s and O 1s photoemission line can be used to derive additional information on the chemisorption bond.

The second example concerns well-ordered adsorbate layers on single crystal surfaces. A wide variety of such layers has been investigated with metals and metal oxides as substrates. The three most important questions one tries to answer in these studies are the following:

1. What is the crystallographic structure of the adsorbate overlayer?
2. What are the adsorption sites (on top of a surface atom, between two or more, etc.)?
3. How are the individual molecules adsorbed (standing, lying, tilted)?

The first question is usually answered by LEED (low-energy electron diffraction) but for the other two ARPS is the method of choice. Figure 29 shows the result of such a study: a dense monolayer of CO on a Ni(110) surface. The layer has p2mg symmetry. Each CO molecule is bound in a bridging position between two neighboring Ni atoms along the $[1\bar{1}0]$ direction. The C-terminus points down to the Ni surface. The CO molecules are tilted by $(17 \pm 2)^\circ$ along the [001] direction in order to avoid the close intermolecular contact along $[1\bar{1}0]$, which would occur if the molecules were oriented perpendicular to the surface. Most of these details resulted from the determination of a two-dimensional band structure with the methods outlined in Section IV.B and its comparison to theoretical results.

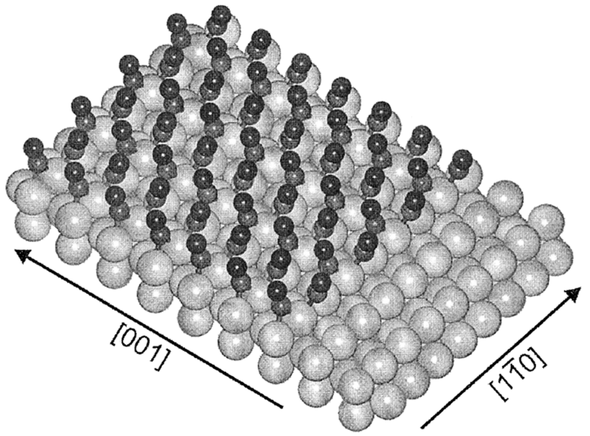


FIGURE 29 Schematic representation of CO adsorbed in a (2×1) p2mg overlayer on a Ni (110) surface. [From Freund, H.-J., and Kuhlenbeck, H. (1995). *In* "Applications of Synchrotron Radiation" (Eberhardt, W., ed.), Springer-Verlag, Berlin, Heidelberg.]

E. Studies in Corrosion

Corrosion is of tremendous economic importance. In the United States, for example, a total damage of more than $300 billion is caused every year by corrosion. Among the important corrosive reactions are all kinds of oxidation reactions, anodic dissolution, passivation, breakdown of passivation, and absorption of hydrogen, usually the first step of hydrogen embrittlement. All these reactions take place in the outermost layers of the material, but many of them are not well understood on an atomic or molecular level. The surface science approach to corrosion tries to clarify the mechanisms of corrosion reactions on that level and to link this understanding to the macroscopic manifestations of corrosion. PES as a highly surface sensitive technique became an important tool in this area of research.

In most cases, the surface of a metal or an alloy, and often also of a polymer, differs from the bulk. This is due to spontaneous reactions with the environment like oxidation, which can lead either to passivation or to ongoing destruction. In alloys, which account for the majority of technically important materials, the composition of the surface often differs from the bulk composition to dealloying or to segregation processes that accompany surface reactions.

The surface zone, which is important for passivation or corrosion, often consists of only a few atomic layers. In these cases, PES can contribute significantly to the investigation of the overlayer. The key is that for most elements, oxides are easily distinguished from the element itself. Angular dependent measurements do not only allow a determination of the thickness of the oxide layer: because of the element specificity of PES, such measurements also can be used to study the change of composition perpendicular to the surface.

Figure 30 shows an example. The sample is an iron/chromium alloy that had been exposed to an aqueous solution of sulfuric acid. The alloy is covered by a surface layer of approximately one nanometer that consists of oxides and hydroxides of iron and chromium. The composition of the material immediately underneath the passive film equals the bulk composition. This is not always the case, as elements are often depleted underneath the passive film if an enrichment of these elements occurs in the film. Technical problems with this kind of investigation mainly concern the transfer of the corroded sample into the vacuum of the spectrometer. Passive layers that contain hydroxides in solution especially tend to degrade in the vacuum.

As mentioned at the beginning of this section, these few examples can only provide a glimpse of a very broad field.

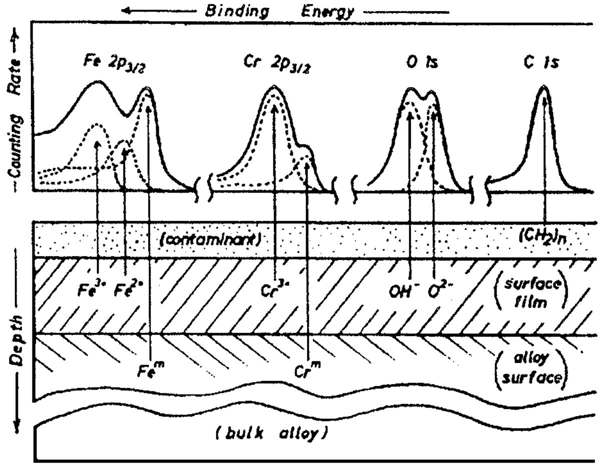


FIGURE 30 Schematic representation of an Fe-Cr alloy surface and the observed XPS spectrum. [From Asami, K., and Hashimoto, K. (1987). *Langmuir* **3,** 897.]

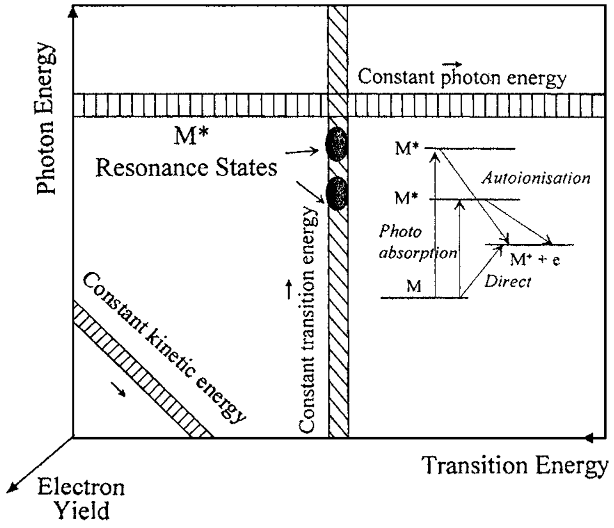


FIGURE 31 Schematic 2DPES, in which the electron yield is displayed as a function of photon and electron energy, illustrating the appearance of features typical for a 2DPES. [From, Sokell, E., *et al.* (1998). *J. Electron Spectrosc.* **94,** 107.]

We hope, however, that they can help the reader to gain an idea of how and where PES can be applied.

V. NEW DIRECTIONS

Since the first edition of the encyclopedia was published, PES has developed quite rapidly in terms of the depths of information that can be obtained from this method, as well as in the range of application. One of the reasons for this development is the growing use of synchrotron radiation for PE work. The development of two-dimensional photoelectron spectroscopy (2DPES), for example, was only possible with the availability of synchrotron radiation. It is a powerful experimental technique, particularly well suited to the comprehensive coverage of the large number of final ion states available in molecular targets and also for the study of the evolution of the effects of postcollision interactions (PCI).

In 2DPES studies surfaces, which indicate electron yield as a function of both electron and photon energy, are derived by measuring many conventional PE spectra. Many 2DPES studies have been concerned with the decay routes of autoionizing photoexcited resonance states. Excited states of the neutral system that are energetically higher than the first ionization potential can decay to lower lying ion states by emission of an electron (autoionization, route 2 in Fig. 2). In many cases, these final ionic states are also directly accessible. A schematic 2DPES of this kind is shown in Fig. 31. The binding energy, which is referred to as transition energy in this plot, is the difference between the incident photon energy and the measured electron kinetic energy. Particular final ion states appear as bands of constant transition energy parallel to the photon energy axis. The intensity of these bands, which results from direct photoionization, is enhanced at certain photon energies by decays from autoionizing resonance states. These exited neutral states appear at well-defined incident energies.

The future application of 2DPES at higher resolution, which is now available at third-generation synchrotron radiation sources, will, among others, permit the study of more closely spaced Rydberg states. Instrumental effects and spectral features arising from impurities, which may be hard to identify in one-dimensional measurements, are readily visible in 2DPES. In order to compare two-dimensional spectra with other work, it is often necessary to extract one-dimensional spectra. This is possible, for example, with a spectrum of constant transition energy, which indicates the population of particular ion states as a function of photon energy.

Photoelectron emission microscopy (PEEM) is another new method which allows the imaging of surfaces while keeping the element sensitivity typical for PE and Auger spectroscopy. Technically, PEEM is a derivative of low-energy electron microscopy (LEEM), an imaging method in which the electrons backscattered from a low-energy electron beam are used to create a magnified image of a surface. The magnification process is similar to the one used in field emission microscopy. In PEEM the electrons that are used for the imaging processes are photoelectrons. Combining the magnification technique used in LEEM with an energy discriminating detector allows one to use only those electrons for the image that originate from a certain valence band region or a certain core line. For example, because of the element specificity of the

core line, this procedure allows the imaging of the lateral distribution of a single element within a multicomponent surface. Important areas of research that may benefit from these kinds of experiments include studies of impurity segregation at grain boundaries, Schottky barrier formation on semiconductors, and diffusion. PEEM is also expected to gain importance in studies of the complex surface structures of highly integrated circuits.

There are two new fields of PES research that have gained increasing interest during the last decade: Clusters and high T_c superconductors. A cluster consists of n atoms or molecules that stick together to form a kind of supermolecule. Progress in preparation techniques now allows the study of ionic and, more recently, neutral clusters with a well-defined number of atoms or molecules. Investigation of a series of such clusters with increasing n yields information on the development of a given physical property from the single atom or molecule to the solid state. PES is especially suited for examining the development of the outermost orbitals of the free atom or free molecule into the valence band of the corresponding solid. Pronounced stabilities for certain values of n are usually found in small clusters. PES has become extremely helpful in explaining these stabilities. For elements that are metals in the solid state, it is of special interest to determine the number of atoms that are necessary to develop metallic properties. In the case of mercury, for example, this problem was successfully studied by PES.

The discovery of the so-called high T_c superconductors with critical temperatures up to 100 K (such as $YBaCu_3O_{6+x}$ and $La_{2-x}Sr_xCuO_4$ and $Bi_2Sr_2Ca_{x-1}Y_xCu_2O_8$) immediately prompted a variety of PES investigations. In the meantime, electron spectroscopy turned out to be the key technique to understand the electronic structure of these cuprates. These experiments are, however, extremely difficult because of the surface sensitivity of PES and the fact that the surface of these materials usually consists of a few layers of nonconducting material, which, due to segregation phenomena, exhibit a chemical constitution different from the bulk. The most reliable experiments have been performed on the Bi-based materials that have a natural cleavage plane with an unreconstructed surface. The electronic structure derived from the PES studies revealed the importance of the CuO_2 planes which appear in all these materials. The compounds can be described as doped Mott insulators. The highest occupied valence band in the insulating parent compounds have in-plane O $2p$ character, so that the systems are charge transfer insulators in the so-called Zaanen–Sawatzky–Allen classification scheme. Further important information came from ARPES. At present the most pertinent questions concern in-plane anisotropies related to the unusual normal state properties of the doped cuprates. Recent experiments on the Bi-based materials indicated that a description of the normal state as a Fermi-liquid with well-defined quasiparticles is not adequate. In contrast, below T_c such well-defined quasiparticles appear to exist.

Further PES investigations have been devoted to the technically important question of corrosion stability. Wet air and CO_2 lead to an increase of hydroxides or carbonates and a depletion of Cu in the surface layers. The depth of the corroded, nonsuperconducting surface zone ranges from a few nanometers up to several tens of nanometers, depending on the nature of the material and on the manner in which the sample is prepared (polycrystalline material, single crystals, or epitactically grown films).

SEE ALSO THE FOLLOWING ARTICLES

Adsorption • Atomic Spectrometry • Auger Electron Spectroscopy • Microwave Molecular Spectroscopy • Mössbauer Spectroscopy • Multiphoton Spectroscopy • Radiation Sources • Surface Chemistry • Vacuum Technology • X-Ray Analysis • X-Ray, Synchrotron Radiation, and Neutron Diffraction

BIBLIOGRAPHY

Asami, K., and Hashimoto, K. (1987). *Langmuir* **3,** 897.

Baalmann, A., *et al.* (1985). *Solid State Commun.* **54,** 583.

BESSY II (1986). "Eine optimierte Undulator/Wiggler-Speicherring Lichtquelle für den VUV- und XUV-Spektralbereich," p. 20, BESSY, Berlin.

Bischof, P., *et al.* (1969). *Helv. Chim. Acta* **52,** 1745.

Briggs, D., and Seah, M. P., eds. (1983). "Practical Surface Analysis by Auger and X-Ray Photoelectron Spectroscopy," Wiley, New York.

Cardona, M., and Ley, L., eds. (1978). "Topics in Applied Physics," Vol. 26, "Photoemission in Solids I," p. 193, Springer-Verlag, Berlin.

Eberhardt, W., ed. (1995). "Application of Synchrotron Radiation," Springer-Verlag, Berlin, Heidelberg.

Eckardt, H., Fritsche, L., and Noffke, J. (1983). *J. Phys. F* **14,** 97.

Eland, J. H. D. (1984). "Photoelectron Spectroscopy," 2nd ed. Butterworths, London.

Freund, H.-J., and Kuhlenbeck, H. (1995). *In* "Applications of Synchrotron Radiation" (W. Eberhardt, ed.), Springer-Verlag, Berlin, Heidelberg.

Freund, H.-J., and Plummer, E. W. (1981). *Phys. Rev. B* **23**, 4859.

Heimann, P., Miusga, H., and Neddermeyer, H. (1979). *Solid State Commun.* **29,** 463.

ICES (1997). "Proceedings of the 7th International Conference on Electron Spectroscopy," ICES 7, Chiba, Japan, September 8–12, Elsevier.

Jung, C., and Bressler, P. R. (1996). *Electron Spectrosc.* **78,** 503.

Kevan, S. D., ed. (1992). "Angle Resolved Photoemission, Theory and Current Applications," Elsevier.

Plummer, E. W., *et al.* (1977). *Phys. Rev. A* **15,** 2339.

Siegbahn, K. (1974). *J. Electron Spectrosc.* **5,** 34.

Siegbahn, K., *et al.* (1971). "ESCA Applied to Free Molecules," Elsevier, Amsterdam.
Sodhi, R. N., and Cavell, R. G. (1983). *J. Electron Spectrosc.* **32,** 283.
Sokell, E., *et al.* (1998). *J. Electron Spectrosc.* **94,** 107.
Vickerman, J. C., ed. (1997). "Surface Analysis—The Principal Techniques," Wiley, Chichester.
Watts, J. F. (1994). "X-ray photoelectron spectroscopy," *Vacuum* **45** (6/7), 653.
Wagner, C. D., *et al.* (1979). "Handbook of XPS," Perkin-Elmer Corp., Eden Prairie, MN.

Raman Spectroscopy

R. P. Van Duyne
C. L. Haynes
Northwestern University

I. Classical Model of Raman Scattering
II. Quantum Mechanical Model of Raman Scattering
III. Appearance of a Raman Spectrum and Units
IV. Polarization Properties of Raman Scattering
V. Basic Experimental Apparatus
VI. Basic Experimental Techniques
VII. Nonlinear Raman Effects
VIII. Chemical and Biochemical Applications
IX. The Solid State
X. Recent Developments—Single Molecule SERS (SMSERS)

GLOSSARY

Anti-Stokes Raman scattering Light scattering in which the photons gain energy as a result of photon–molecule collisions.

Chromophore Molecule or part of a molecule that absorbs light.

Coherent anti-Stokes Raman spectroscopy (CARS) Spectroscopic technique based on a nonlinear event in which two cofocused laser beams give rise to coherent Raman scattering.

Depolarization ratio Ratio of intensities of light scattered perpendicular and parallel to the electric field vector ($\boldsymbol{E}$) of the incident radiation.

Electric polarizability Proportionality factor relating the induced dipole moment to the electric field strength of a light wave.

Induced dipole Dipole induced in a system as a result of the electric field of a light wave.

Localized Surface Plasmon Resonance (LSPR) Electromagnetic wave localized on a nanometer-scale roughened metal surface when an incoming photon strikes a resonance condition and causes a collective oscillation of electrons.

Nonlinear Raman effects Raman-like processes occurring at high photon densities, in which second- and higher-order terms in the expression for the induced dipole $\mu_{\mathbf{ind}} = \alpha \boldsymbol{E} + 1/2\beta \boldsymbol{E} \cdot \boldsymbol{E} + \cdots$, must be taken into account.

Polarizability tensor The 3×3 tensor α in the relation

$\mu_{\mathbf{ind}} = \alpha \boldsymbol{E}$ between the induced dipole $\mu_{\mathbf{ind}}$ vector and the electric field $\boldsymbol{E}$ in linear Raman scattering. $\alpha_{\rho\sigma}$ is the polarization element where ρ and σ are the Cartesian coordinates x, y, or z.

Raman microscope Combination of microscope optics and a Raman spectrometer equipped with a charge-coupled device camera that enables the observer to view a microscopic object according to its Raman-active vibrations.

Resonance Raman scattering The intensity of Raman scattering is amplified by 10^3–10^5 times when the wavelength of the incident light is coincident with an intense molecular electronic absorption transition.

SPP-Raman The intensity of Raman scattering is enhanced by 10^2–10^3 times when a smooth Ag or Au surface is excited by incoming radiation.

Stokes Raman scattering Light scattering in which the photons lose energy as a result of photon–molecule collisions.

Surface-enhanced hyper-Raman scattering (SEHRS) The intensity of the two-photon Raman scattering is enhanced by 10^{11}–10^{12} times from molecules (i) adsorbed at a Ag, Cu, or Au surface roughened on the nanometer scale; and (ii) having an electronic absorption band at or near the excitation wavelength.

Surface-enhanced Raman scattering (SERS) The intensity of Raman scattering is amplified by 10^6–10^8 times from molecules adsorbed at a Ag, Cu, or Au surface roughened on the nanometer scale.

Surface-enhanced resonance Raman scattering (SERRS) The intensity of Raman scattering is amplified by 10^9–10^{11} times from molecules adsorbed at a Ag, Cu, or Au surface roughened on the nanometer scale.

Surface Plasmon Polariton (SPP) Evanescent, p-polarized, nonradiative electromagnetic wave associated with the propagation of a charge wave at a smooth metal–dielectric interface.

Time-resolved resonance Raman spectroscopy (TR^{3}S) Study of transient chemical species or excited electronic states via their Raman spectra, often with pulsed lasers and multichannel photon detectors.

Ultraviolet resonance Raman scattering (UVRRS) Selective excitation of a UV resonance enhanced Raman-active mode of vibration.

RAMAN SPECTROSCOPY is a technique that uses scattered light resulting from photon–molecule collisions to investigate molecular properties. When a monochromatic light beam is incident on systems such as transparent gases, liquids, or solids, most of it is transmitted without change. However, a very small portion of the incident light is scattered. Although most of the scattered light has the same wavelength as the incident radiation, a small part of it occurs at different wavelengths. The scattering of light at different wavelengths is called Raman scattering, after the Indian scientist Sir C. V. Raman, who, with K. S. Krishnan, first reported the phenomenon in liquids in 1928. In the same year Landsberg and Mandelstam in Russia reported a change in frequency of the light scattered from quartz. The physical origin of Raman scattering lies in inelastic collisions between the molecules composing the system (e.g., the liquid) and photons, the particles composing the light beam. "Inelastic collision" means that there is an exchange of energy between the photon and the molecule and a consequent change in energy, and hence wavelength, of the photon. Moreover, since total energy is conserved during the scattering process, the energy gained or lost by the photon must equal the energy change within the molecule. By measuring the energy gained or lost by the photon, changes in molecular energy can be probed. The changes are usually related to rotational, electronic, and vibrational energy levels. Since much of present-day Raman spectroscopy is concerned with vibrational energy levels, thisarticle will concentrate on vibrational properties. However, the theoretical principles and experimental techniques enunciated for the vibrational states are readily extended to other types of molecular properties.

I. CLASSICAL MODEL OF RAMAN SCATTERING

Light is a traveling wave of electric and magnetic fields, of which only the electric component gives rise to Raman scattering. When a light wave meets a molecule consisting of electrons and nuclei, the electric field of the wave at any instant will be the same throughout the molecule because the molecule, perhaps 1 nm in size, is small compared to the wavelength of the light, which is typically 500 nm. Thus, the field will exert the same force on all electrons in the molecule and will tend to displace them from their average positions around the positively charged nuclei. For the Raman process, the displacements result in an induced dipole moment, $\mu_{\mathbf{ind}}$ in the molecule that is, to a good approximation, proportional to the electric field strength $\boldsymbol{E}$.

Thus

$$\mu_{\mathbf{ind}} = \alpha \boldsymbol{E}, \tag{1}$$

where the proportionality factor α is called the electric polarizability of the molecule. In general, the vector $\mu_{\mathbf{ind}}$ will have a different direction from the vector $\boldsymbol{E}$, and therefore, α is not a simple scalar quantity. In fact, the magnitudes of the three components defining the induced dipole moment,

namely μ_x, μ_y, and μ_z, are related to the magnitudes of the electric field $\boldsymbol{E}$ by the three relations

$$\begin{aligned} \mu_x &= \alpha_{xx}\boldsymbol{E}_x + \alpha_{xy}\boldsymbol{E}_y + \alpha_{xz}\boldsymbol{E}_z \\ \mu_y &= \alpha_{yx}\boldsymbol{E}_x + \alpha_{yy}\boldsymbol{E}_y + \alpha_{yz}\boldsymbol{E}_z \\ \mu_z &= \alpha_{zx}\boldsymbol{E}_x + \alpha_{zy}\boldsymbol{E}_y + \alpha_{zz}\boldsymbol{E}_z. \end{aligned} \tag{2}$$

The equations of (2) express the fact that all three components of $\boldsymbol{E}$ make contributions to each of three components of $\boldsymbol{\mu}_{\mathbf{ind}}$. The nine coefficients α_{ij} are called the components of the polarizability with the subscript i denoting the component of $\boldsymbol{\mu}_{\mathbf{ind}}$ and the subscripts j denoting the component of $\boldsymbol{E}$ related by the element α_{ij}. As a consequence of Eq. (2), α is said to be a tensor.

The electric field of the light wave varies with time. If a fixed molecule is irradiated with monochromatic radiation of frequency ν_0, expressed in hertz (cycles per second), which is plane polarized in the z direction, $\boldsymbol{E}_z$ as a function of time is given by

$$\boldsymbol{E}_z(t) = \boldsymbol{E}_{\max}\cos 2\pi\nu_0 t, \tag{3}$$

where $\boldsymbol{E}_{\max}$ is the value of $\boldsymbol{E}_z$ at its maximum and t is the time in seconds from an arbitrary starting time. Thus, for the z component of μ_{ind}

$$\mu_z(t) = \alpha_{zz}\boldsymbol{E}_{\max}\cos 2\pi\nu_0 t. \tag{4}$$

Since μ_z depends on α_{zz}, as well as $\boldsymbol{E}_z$, the properties of the molecule can change μ_z via α. In the present context α varies with time. As a consequence of the vibrations of the molecule, the ease with which electrons may be displaced by the electric field depends on how tightly they are bound to the nuclei, which in turn depends on the internuclear separation. The result of the time dependence of α and $\boldsymbol{E}$ on $\boldsymbol{\mu}_{\mathbf{ind}}$ can be seen by considering the simple example of a single diatomic molecule.

For a diatomic molecule, the difference from equilibrium in the internuclear distance at time t, $\Delta r(t)$, can be rewritten

$$\Delta r(t) = \Delta r_{\max}\cos(2\pi\nu_{\text{vib}}t + \phi), \tag{5}$$

where ν_{vib} is the vibrational frequency of the molecule in hertz and $\Delta r_{\max}$ is the maximum extension of the distance between the two atoms. By designating time t to have the same starting point as the time scale for the light wave in Eq. (3), the phase constant ϕ may be equated to zero.

Using the postulate that the polarizability of the diatomic molecule depends linearly on Δr, the α_{zz} component of α can be written

$$\begin{aligned} \alpha_{zz}(t) &= \alpha_{zz}^{\text{equil}} + \frac{d\alpha_{zz}}{dr}\Delta r(t) \\ &= \alpha_{zz}^{\text{equil}} + \frac{d\alpha_{zz}}{dr}\Delta r_{\max}\cos 2\pi\nu_{\text{vib}}t. \end{aligned} \tag{6}$$

The constant $\alpha_{zz}^{\text{equil}}$ is the polarizability element of the nonvibrating molecule and $d\alpha_{zz}/dr$ characterizes the manner in which the polarizability changes with r.

Substituting the expressions for α_{zz} into Eq. (4), the dependence of μ_{ind} on the time fluctuations of both α and $\boldsymbol{E}$ becomes

$$\begin{aligned} \mu_z(t) &= \alpha_{zz}^{\text{equil}}\boldsymbol{E}_{\max}\cos 2\pi\nu_0 t \\ &\quad + \frac{d\alpha_{zz}}{dr}\Delta r_{\max}\boldsymbol{E}_{\max}\cos 2\pi\nu_0 t\cos 2\pi\nu_{\text{vib}}t. \end{aligned} \tag{7}$$

Using the identity $\cos\theta\cos\phi = 1/2[\cos(\theta+\phi)+\cos(\theta-\phi)]$, Eq. (7) becomes

$$\begin{aligned} \mu_z(t) &= \alpha_{zz}^{\text{equil}}\boldsymbol{E}_{\max}\cos 2\pi\nu_0 t \\ &\quad + \frac{1}{2}\frac{d\alpha_{zz}}{dr}\Delta r_{\max}\boldsymbol{E}_{\max}\cos 2\pi(\nu_0 + \nu_{\text{vib}})t \\ &\quad + \frac{1}{2}\frac{d\alpha_{zz}}{dr}\Delta r_{\max}\boldsymbol{E}_{\max}\cos 2\pi(\nu_0 - \nu_{\text{vib}})t. \end{aligned} \tag{8}$$

Equation (8) demonstrates that when a light wave interacts with a vibrating diatomic molecule, the induced dipole moment, in this case exemplified by μ_z, has three components contributing to its time dependence. The first term on the right of Eq. (8) is a component vibrating with the frequency of the incident light and with a magnitude determined by $\alpha_{zz}^{\text{equil}}$ and $\boldsymbol{E}_{\max}$. According to classical electromagnetic theory, an oscillating dipole radiates energy in the form of scattered light. Thus, as a result of the first term in Eq. (8), light of the incident frequency ν_0 will be emitted and will be observed in directions different from that of the incoming light. This is the phenomenon of Rayleigh scattering. The second term is a component vibrating at a frequency that is the sum of the frequencies of the light and the molecular vibration. The scattered light arising from this second term is known as anti-Stokes Raman scattering. The third term is a component vibrating at a frequency given by that of the light wave minus that of the molecule, and the scattered light resulting from this term is known is Stokes Raman scattering. Both these components have magnitudes that depend on the field strength of the light, the amplitude of vibration, and the polarizability derivative $d\alpha_{zz}/dr$. The appearance of scattered radiation, for instance, at $\nu_0 + \nu_{\text{vib}}$ arising from the second term in Eq. (8) or $\nu_0 - \nu_{\text{vib}}$ from the third term in Eq. (8) means that by analyzing the scattered light, we can monitor the vibrations within a molecule. It is this ability to measure molecular vibrations (and electronic or rotational

transitions) that gives the Raman effect its importance in the study of molecules.

The information about molecular vibrational frequencies provided by infrared absorption spectroscopy is of the same kind as that provided by the ν_{vib} values of the Raman lines. Moreover, for molecules with little or no symmetry, the infrared and Raman spectra can, in some cases, have quite similar appearances. However, information from both techniques is important for molecules having a center of symmetry, since there can be complementarity between the Raman and infrared spectra with no correspondence between the main features exhibited in each spectrum. The complementarity derives from the rule of mutual exclusion, which states that for molecules with a center of symmetry, vibrational transitions that are allowed in the infrared are forbidden in the Raman effect and vice versa. The reason for the complementarity of Raman and infrared spectra lies in the different natures of the physical processes involved in the two effects. The Raman process is a scattering effect involving an induced dipole $\boldsymbol{\mu}_{\mathbf{ind}}$, which in turn depends on a change in molecular polarizability during a vibration. In contrast, infrared spectroscopy is an absorption process caused by a change in the permanent molecular dipole μ with change in bond length during a vibration.

II. QUANTUM MECHANICAL MODEL OF RAMAN SCATTERING

The quantum mechanical approach to the scattering process is quite different from the classic model: the wave–particle duality of a light beam is incorporated by considering that the beam is made up of packets or quanta of light particles known as photons. Moreover, the quantization of molecular energy levels is taken into account and a means is provided for calculating the polarizability α, and thus Raman intensities, in terms of the electronic properties of a molecule.

It is a reasonable approximation, for a gaseous diatomic molecule, to write the molecular energy $\boldsymbol{E}_{\mathrm{max}}$ as a sum of terms:

$$\boldsymbol{E}_{\mathrm{max}} = \boldsymbol{E}_{\mathrm{elec}} + \boldsymbol{E}_{\mathrm{vib}}, \qquad (9)$$

where the respective subscripts refer to electronic and vibrational components of the total energy. For present purposes the contributions due to molecular rotation and translation can be ignored. Electronic energy transitions involve much larger quantities of energy than vibrational transitions,with values of 10,000 to 50,000 cm^{-1} for the former and 10 to 4000 cm^{-1} for the latter. This situation is depicted in Fig. 1 with a large energy spacing between the ground and excited electronic states and smaller spacings between the vibrational levels contained within each state. Moreover, for the vibrational energy $\boldsymbol{E}_{\mathrm{vib}}$ of a diatomic molecule it is a good approximation to write

$$\boldsymbol{E}_{\mathrm{vib}} = \left(v + \tfrac{1}{2}\right) h\nu_{\mathrm{vib}} \qquad v = 0, 1, 2, \ldots, \qquad (10)$$

where the vibrational quantum number v has only integral values, so the vibrational energy levels in the ground state in Fig. 1 are equally spaced by the amount $h\nu_{\mathrm{vib}}$.

In the quantum mechanical model, light scattering is depicted as a two-photon process. The first step in this process is the combination of a photon and a molecule to raise the molecule to a higher-energy, short-lived state. This state is reached by the upward arrows in Fig. 1 and, as shown in the figure, may or may not correspond to a quantized energy state of the molecule. The second step, indicated by the downward arrows in Fig. 1, involves the release of a photon after a very short time interval ($<10^{-14}$ sec). The energy of this second photon is given by the length of the downward arrows in Fig. 1. For Rayleigh scattering the upward and downward transitions have the same length and have therefore, apart from a change in sign, the same energies. Thus, in the Rayleigh process, no change in frequency of the photon occurs.

The various kinds of Raman processes can now be outlined. If the downward arrow stops on a vibrational energy level that is higher than the starting level, a Stokes process has occurred. In this case, the second photon has a frequency $\nu_0 - \nu_{\mathrm{vib}}$ corresponding to the third term in Eq. (8) of the discussion of the classic model for the Raman effect. Conversely, an anti-Stokes process results from the transition terminating at a vibrational energy level lower than the starting level. In the anti-Stokes process, the second photon has a frequency $\nu_0 + \nu_{\mathrm{vib}}$, giving the same result as the second term in Eq. (8). Of course, in both processes total energy is conserved, so for Stokes scattering the molecule gains a quantum of energy, $h\nu_{\mathrm{vib}}$, while for anti-Stokes scattering the reverse is true. For both Stokes and anti-Stokes processes a selection rule can be derived from Eq. (10) which says that v'', in Fig. 1, can only change by ± 1. Thus Eq. (8) derived from the classic model agrees with the results obtained by considering quantized energy levels in that both models predict that the difference in frequency between the incident and scattered light, $\nu_0 - (\nu_0 - \nu_{\mathrm{vib}})$, corresponds directly to the molecular vibrational frequency ν_{vib}. The quantum mechanical model also illustrates an important generalization, namely, that the position of Raman peaks is a property solely of the electronic ground state. This follows from the fact that $h\nu_{\mathrm{vib}}$ is the energy change in a vibrational transition within the lower or ground electronic state in Fig. 1.

In the classical model Eq. (8) indicates no difference in the expected intensities of Stokes and anti-Stokes

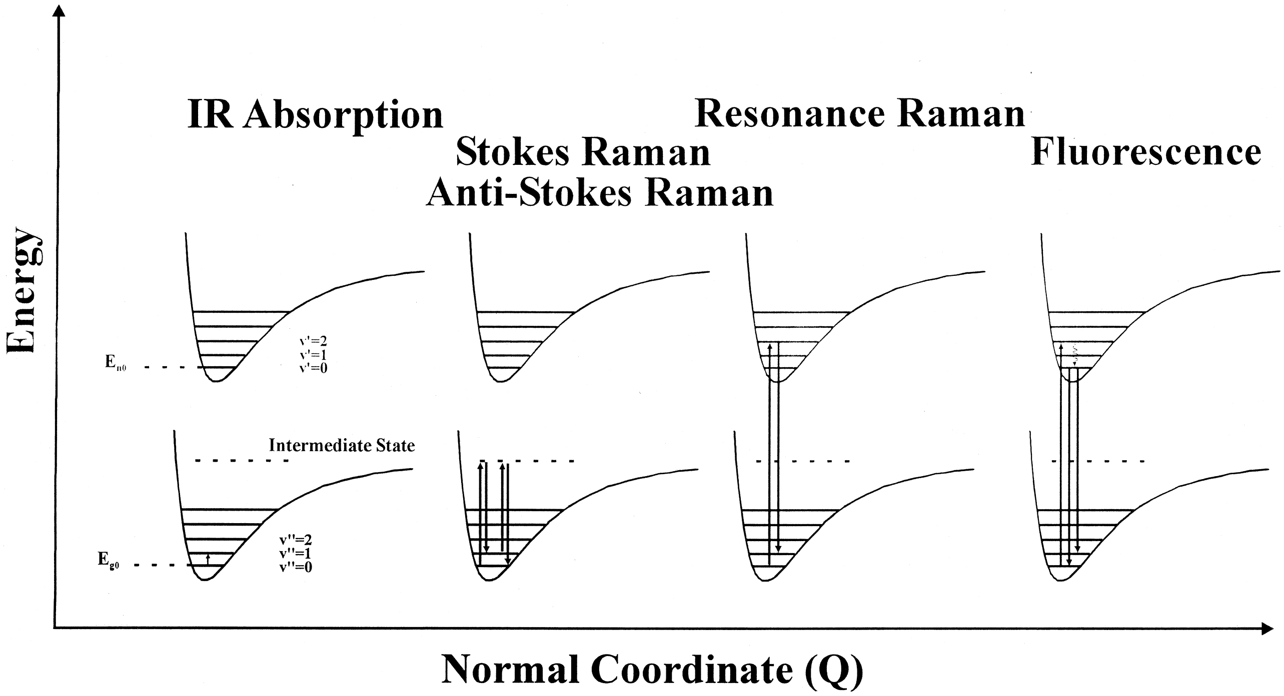


FIGURE 1 Some of the possible consequences of a photon–molecule interaction. The lengths of the upward-pointing arrows are proportional to the frequencies of the incoming light, and the lengths of the downward-pointing arrows are proportional to the frequency to the scattered (or in the case of fluresence, emitted) light. The energy spacing between the ν levels is proportional to ν_{vib} ($\Delta E = h\nu_{vib}$).

transitions, since the coefficients of the two terms in the equation are the same.

However, the model of quantized energy levels depicted in Fig. 1 shows that for anti-Stokes transitions to take place the molecule must be in a higher ($\nu'' > 0$) vibrational state within the electronic ground state. Since the population of vibrational levels is governed by a Boltzmann distribution, only a small percentage of molecules are in higher vibrational states. The ratio of the numbers of molecules in the $\nu'' = 1$ and $\nu'' = 0$ vibrational states in the ground electronic state (Fig. 1), which we will denote N_1 and N_0, respectively, is given by

$$\frac{N_1}{N_0} = \exp\left(-\frac{h\nu_{\mathrm{vib}}}{kT}\right), \tag{11}$$

where h is Planck's constant, T the absolute temperature, and k the Boltzmann constant. When $T = 300$ K and the vibrational frequency is 480 cm^{-1}, N_1 is $0.1N_0$. As a result of the exponential nature of Eq. (11) an anti-Stokes line at 3×480 cm^{-1}, or 1440 cm^{-1}, would be 0.001 times as strong as the corresponding Stokes line. In practice this means that the feeble anti-Stokes scattering is usually ignored in conventional Raman spectroscopy and only the Stokes spectrum is recorded.

In normal Raman scattering, the energy of the incident light is considerably less than that needed to reach the higher energy electronic state shown in Fig. 1. Consider the situation when the photon energy is approaching that of the energy gap between the lower and higher electronic states. A preresonance Raman transition is due to a light frequency that has almost enough energy to produce direct electronic absorption by the molecule. Under this condition the intensity of Raman scattering shows a marked increase. For normal or nonresonant conditions Raman intensities are proportional to the fourth power of the scattered light frequency ν_s. However, as preresonance Raman conditions are approached the intensity of scattering goes up much more rapidly than ν_s^4.

A slight increase in the energy of the exciting radiation over that for the preresonance case will place the upward transition in Fig. 1 within the higher electronic state. Absorption of a photon can now occur and by the prompt reemission of a second photon can give rise to the resonance Raman process. Band intensities in resonance Raman spectra can be orders of magnitude greater than those in normal Raman spectra. The reason for this can be seen by considering the quantum mechanical expression for Raman intensities.

In the quantum mechanical treatment for randomly oriented molecules, the total intensity of the scattered light resulting from a molecular transition between states m and n is

$$I = K(\nu_0 \pm \nu_{mn})^4 I_0 \sum_{\rho\sigma} |(\alpha_{\rho\sigma})_{mn}|^2, \quad (12)$$

where $|\alpha_{\rho\sigma}|_{mn}$ is the transition polarizability tensor. This quantity can be derived from second-order perturbation theory to give

$$(\alpha_{\rho\sigma})_{nm} = \frac{1}{h}\sum_r \frac{\langle n|\mu_\rho|r\rangle\langle r|\mu_\sigma|m\rangle}{\nu_{rm} - \nu_0 + i\Gamma_r} + \frac{\langle n|\mu_\sigma|r\rangle\langle|\mu_\rho|m\rangle}{\nu_m - \nu_0 + i\Gamma_r}. \quad (13)$$

In these expressions the molecule is considered to be in the molecular state m. It is perturbed by an electromagnetic wave of frequency ν_0 and intensity I_0, causing the transition to a state n and scattering light of frequency $\nu_0 \pm \nu_{mn}$. The sum over index r covers all of the quantum mechanical eigenstates of the molecule, h is Planck's constant, and Γ_r is a damping constant that takes into account the finite lifetime of each molecular state. The $\langle n|\mu_p|r\rangle$ and other terms represent the amplitudes of the electric dipole transition moments, where μ_p is the electric dipole moment operator along direction p. Immediately, we see from Eq. (13) that as $h\nu_0$ approaches the energy of an allowed molecular transition $h\nu_{rm}$, $(\nu_{rm} - \nu_0 + i\Gamma_r)$ becomes small and, consequently, one term in the sum becomes very large. This resonance condition is the origin of resonance Raman scattering (RRS).

An important practical outcome of resonance Raman scattering is that the accompanying intensity enhancement allows one to obtain the Raman spectrum of molecules with suitable electronic absorption bands at low concentration, for instance, in the 10^{-5} M range in solution. At the same time, resonance Raman scattering enables the spectroscopist to selectively and specifically "pick out" the Raman spectrum of an absorbing molecule in a complex environment that contributes only a weak normal Raman spectrum. This property is used to great effect in biological Raman studies, since the chromophore is often found at the site of biological activity.

III. APPEARANCE OF A RAMAN SPECTRUM AND UNITS

The appearance of a vibrational Raman spectrum is illustrated by the spectrum of benzene shown in Fig. 2. Analysis of the scattered light perpendicular to the direction of the beam yields the Raman spectrum (see Section V). The most readily discernible features are the ring breathing mode at 992 cm^{-1} and the aromatic C–H stretch mode at 3063 cm^{-1}. These Raman signals arise from an exchange of energy between the incoming photons and the corresponding vibrational motion of the benzene molecules.

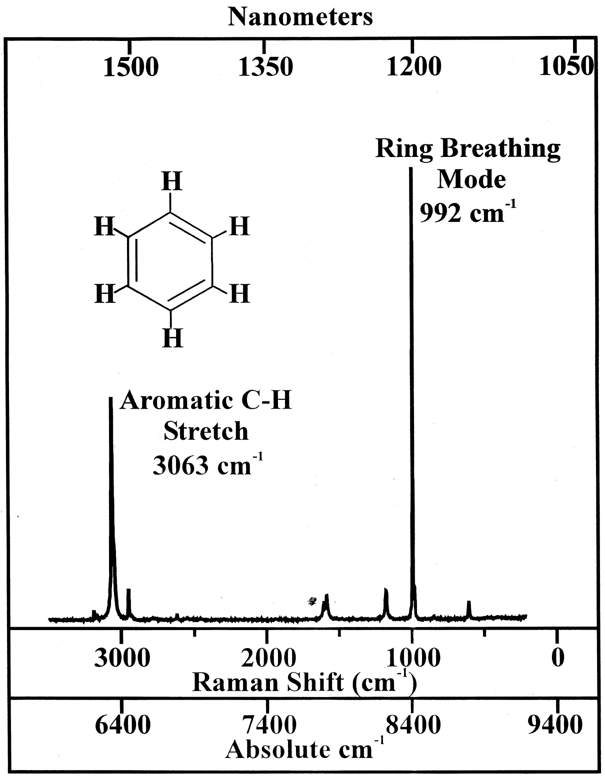


FIGURE 2 FT-Raman spectrum of neat benzene.

In Fig. 2 the abscissa represents an energy difference between the scattered and incoming photons with the zero-point energy difference being at the wavelength of excitation. It is therefore apparent that more energy is required to bring about the aromatic stretching motion than the ring breathing motion. It is also apparent from the intensities of the peaks in the spectrum that the exchange of energy between molecules and photons giving rise to the ring breathing mode is more probable than that giving rise to the aromatic C–H stretching feature.

A monochromatic light beam is characterized by its wavelength λ, its power, and its polarization. Instead of quoting the wavelength of a light beam, this property is often given in terms of the equivalent units of frequency or wavenumber. These quantities are related to wavelength in the following manner:

$$\nu = c/\lambda \text{ hertz (or cycles per second)}, \quad (14)$$

where ν is frequency, c the speed of light (2.998×10^{10} cm/sec in vacuum), and λ must be expressed in the same units as c.

A wavenumber (often denoted $\bar{\nu}$ or ω)

$$= 1/\lambda \text{ cm}^{-1}, \quad (15)$$

where λ is expressed in centimeters.

The wavelength of light is commonly expressed in angstroms (Å) or nanometers (nm):

$$10\ \text{Å} = 1\ \text{nm} = 10^{-9}\ \text{m}. \tag{16}$$

Finally, the wavelength (or frequency) of light is related to the energy $\boldsymbol{E}$ by

$$\boldsymbol{E} = h\nu, \tag{17}$$

where h is Planck's constant.

The units and relationships between the three axes in Fig. 2 should now be clear. The wavelength of the laser line used to excite the Raman spectrum is 1064 nm, which is equivalent to 9399 cm^{-1}. When discussing energies it is simpler to use wavenumbers since, from Eq. (17), energy is linear with wavenumber (and therefore proportional to 1/wavelength). The exchange of energy between the incoming photons and the vibrational energy transitions giving rise to the ring breathing and aromatic C–H stretch features results in a loss of energy for the photons equivalent to 992 and 3063 cm^{-1} (multiplied by hc), respectively. Thus, on the scale of absolute reciprocal centimeters, the ring breathing and aromatic C–H stretch appear at $9399 - 992 = 8407\ \text{cm}^{-1}$ and $9399 - 3063 = 6336\ \text{cm}^{-1}$, respectively. This can be seen by referring to the bottom scale in Fig. 2. The usual convention in Raman spectroscopy is to quote the positions of the vibrational peaks as the difference between the absolute wavenumbers of the exciting line and the absolute wave numbers of the scattered photon. Thus, the Raman shift scale seen in Fig. 2 is normally the only scale encountered in published Raman spectra. The equivalent wavelength scale in nanometers is shown at the top of Fig 2.

IV. POLARIZATION PROPERTIES OF RAMAN SCATTERING

Most natural light is unpolarized, which in a simplified form means the electric field ($\boldsymbol{E}$) vector performs linear oscillations of constant amplitude in a plane perpendicular to the light path but the orientation of the $\boldsymbol{E}$ vector within the plane is completely random. Light from lasers usually has the special property of being plane polarized; that is, the terminus of $\boldsymbol{E}$ varies periodically within a single plane through the light path. Thus, for the plane-polarized beam shown in Fig. 3 the $\boldsymbol{E}$ vector remains in the XZ plane but varies in magnitude along the light path according to the hatched lines. In Fig. 3 the laser beam is depicted as a vertical line starting at the bottom of the diagram. Information on the scattering process and on the assignments of Raman bands can be gained by analyzing the scattered light parallel and perpendicular to the incoming $\boldsymbol{E}$ vector.

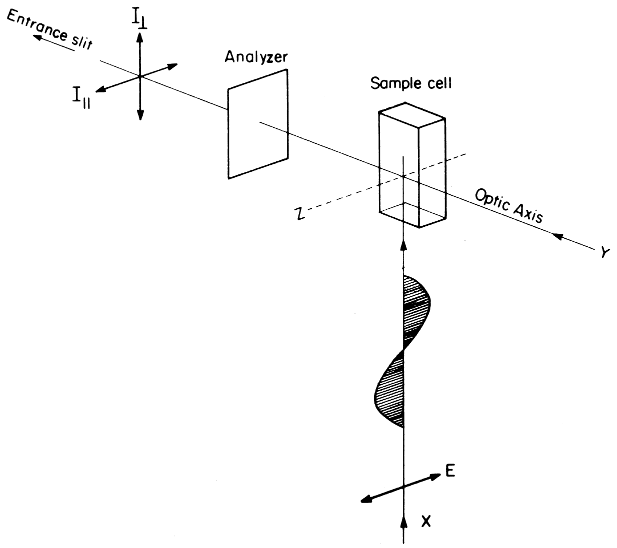


FIGURE 3 Orientation of the $\boldsymbol{E}$ vectors of the incoming ($\boldsymbol{E}$) and scattered light parallel ($I_\parallel$) and perpendicular ($I_\perp$) to the incident $\boldsymbol{E}$ vector. The depolarization ratio is $\rho = I_\perp / I_\parallel$. [Reproduced from Carey, P. R. (1982). "Biochemical Applications of Raman and Resonance Raman Spectroscopies," Academic Press, New York, by permission.]

The *depolarization ratio* ρ of a feature in the Raman spectrum is defined as

$$\rho = I_\perp \mid I_\parallel, \tag{18}$$

where $I_\perp$, and $I_\parallel$, are the intensities of Raman radiation of a given frequency that is polarized, respectively, in a perpendicular and parallel orientation normal to the plane of the incident beam. These relationships are illustrated in Fig. 3. Changes in the polarization of the incident light on scattering result from the tensorial nature of the interaction denoted in Eq. (2). In fact, for a single crystal the individual elements of the scattering tensor a may be related to $I_\perp$ and $I_\parallel$ (see Section VIII.A). For fluids, however, the molecules are randomly oriented with respect to the laboratory-fixed coordinate system used to define α, and when an average over molecular orientations is made, $I_\perp$ and $I_\parallel$ are found to be related to certain combinations of the components of α. These combinations are known as the tensor invariants. Thus, for randomly oriented molecules, when the intensities $I_\perp$ and $I_\parallel$ are computed, they are found to be

$$I_\perp = \text{const}(3g^{\text{s}} + 5g^{\text{a}})\vec{\boldsymbol{E}}_z^2 \tag{19}$$

and

$$I_\parallel = \text{const}(10g^0 + 4g^{\text{s}})\vec{\boldsymbol{E}}_z^2, \tag{20}$$

where as in Fig. 3 the incident light is taken to be polarized in the Z direction, and $\vec{\boldsymbol{E}}_z^2$ is the average value of the square of $\boldsymbol{E}$, to which the intensity of the incident light is

proportional. The depolarization ratio is therefore given by

$$\rho = \frac{3g^{s} + 5g^{a}}{10g^{0} + 4g^{s}}, \qquad (21)$$

where the tensor invariants, known as the isotropic, symmetric anisotropic, and antisymmetric anisotropic components, are

$$\begin{aligned} g^{0} &= \tfrac{1}{3}(\alpha_{xx} + \alpha_{yy} + \alpha_{zz})^{2} \\ g^{s} &= \tfrac{1}{3}\left[(\alpha_{xx} - \alpha_{yy})^{2} + (\alpha_{xx} - \alpha_{zz})^{2} + (\alpha_{yy} - \alpha_{zz})^{2}\right] \\ &\quad + \tfrac{1}{2}\left[(\alpha_{xy} + \alpha_{yx})^{2} + (\alpha_{xz} + \alpha_{zx})^{2} + (\alpha_{yz} + \alpha_{zy})^{2}\right] \\ g^{a} &= \tfrac{1}{2}\left[(\alpha_{xy} - \alpha_{yx})^{2} + (\alpha_{xz} - \alpha_{zx})^{2} + (\alpha_{yz} - \alpha_{zy})^{2}\right]. \end{aligned} \qquad (22)$$

Under all nonresonance conditions, and usually for resonance conditions, $g^{a} = 0$. This being the case,

$$\rho = \frac{3g^{s}}{10g^{0} + 4g^{s}}, \qquad (23)$$

and since in vibrational Raman scattering either g^{s} or g^{0} can be zero, it follows from Eq. (23) that $0 \le \rho \le 3/4$. Detailed calculations show that for normal vibrations that do not preserve molecular symmetry during the motion of the nuclei (nontotally symmetric modes), $g^{0} = 0$ and hence, from Eq. (23), $\rho = 3/4$. For modes that do preserve molecular symmetry neither g^{s} nor g^{0} is zero, leading to values of $\rho \ge 3/4$ for these modes, and, in fact, ρ is often found to be substantially less than 3/4 for totally symmetric vibrations.

For totally symmetric modes, $\rho \le 0.75$
For nontotally symmetric modes, $\rho = 0.75$

Under resonance or near-resonance conditions in certain rare instances, $g^{a} \ne 0$. This arises because, for certain modes, α can become antisymmetric with $\alpha_{\rho\sigma} = -\alpha_{\rho\sigma}$, giving rise to "anomalously" polarized bands with $\rho > 3/4$. In the limiting case when $g^{s} = g^{0} = 0$, but $g^{a} \ne 0$, the phenomenon of inverse polarization occurs with, in theory, $\rho = \infty$. In the usual case when $g^{a} = 0$, the relative values of the two tensor invariants g^{0} and g^{s} can be determined from Eq. (23) by the measurement of ρ. However, for $g^{a} \ne 0$ there are *three* tensor invariants, which, in general, cannot be determined by measurement of ρ alone. To determine these three invariants, it is necessary to use circularly polarized incident light and to analyze the polarizations of the forward or backscattered radiation.

V. BASIC EXPERIMENTAL APPARATUS

A. Radiation Sources

In order to generate a detectable quantity of inelastically scattered photons, it is necessary to have an intense, monochromatic light source. Thus, the advent of the laser was an important event in the development of Raman spectroscopy as a general analytical technique. The lasers most often used in current Raman experiments include Ar^{+}, Kr^{+}, He–Ne, Ti:sapphire, Nd:YAG and diode lasers. Ar^{+} lasers are the most commonly used Raman lasers with irradiation wavelengths of 457.9, 488, and 514.5 nm. Kr^{+} lasers offer wavelengths of 406, 647, and 752 nm. Helium–Neon (He–Ne) lasers operate at a wavelength of 632.8 nm and can only be used for low power applications. Ti:sapphire lasers offer a tunable alternative in the red and near infrared portion of the spectrum. Neodymium-doped yttrium aluminum garnet (Nd:YAG) lasers are pulsed lasers operating at wavelengths of 1064 nm, 532 nm (second harmonic), 355 nm (third harmonic), and 266 nm (fourth harmonic). In recent years, the diode-pumped solid-state laser has permeated the Raman laser market as a Nd:YAG replacement. The advantages of diode lasers compared to the flashlamp-based Nd:YAG include increased reliability, electrical efficiency, less required voltage, less waste heat, and a compact design. Frequency doubled diode pumped Nd:YAG lasers, operating at 532 nm, are a compact, efficient alternative to the Ar^{+} laser.

B. Dispersion Elements

The basis of the dispersing process is depicted in Fig. 4 by a single monochromator. Single monochromators are advantageous because they have higher throughput and collection angles than double and triple monochromators. Before Rayleigh line rejection filters were available (see Section V.D), double and triple monochromators were necessary to separate the Raman photons from the overwhelming number of Rayleigh photons. As shown in Fig. 4, the scattered light enters the monochromator through a single slit and is reflected from a mirror to a fixed diffraction grating. This diffraction grating disperses the scattered light spatially on the basis of frequency, then directs it to another mirror which sends the separated light to a charge-coupled device (CCD) detector. If a photomultiplier tube (PMT) detector is used instead of a CCD detector, the diffraction grating must be slowly turned so that the lines of the spectrum move in succession across the slit and are detected and recorded sequentially.

C. Detectors

Mulitchannel CCD detectors are used almost exclusively in modern Raman spectrometers, replacing PMT and

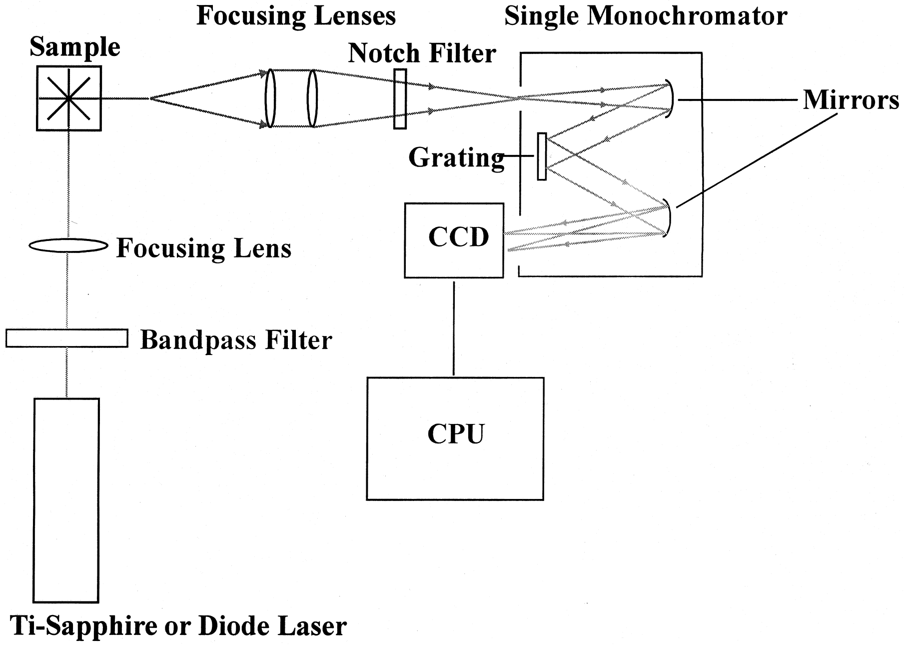


FIGURE 4 Schematic diagram of a Raman spectrometer showing the options of a single monochromator and multichannel detection.

intensified photodiode array (IPDA) detectors. A CCD is a multichannel detector made up of large arrays of individual metal–oxide–silicon capacitors. All the Raman lines are registered on different elements of the detector at the same time. Thus, it is possible to observe the entire Raman spectrum on a television screen or an oscilloscope in real time. When using multichannel detection, the grating is turned only to change the spectral region across the detector.

Recent advances in photodiode technology have produced a new detector for high-sensitivity, low-light detector applications in which PMTs, and even CCDs, are unsuitable. This ultrasensitive device is known as an avalanche photodiode (APD) and is based on standard photodiode technology. The extra sensitivity of the APD is gained because a high bias potential accelerates electrons to ionize semiconductor atoms. These semiconductor atoms trigger the release of free electrons which are then accelerated to continue the cycle in an "avalanche effect." This effect gives rise to thousands of electrons for each initial impinging photon.

D. Filters

As mentioned in previous sections, optical filters play an important role in Raman spectrometers. An interference filter is usually inserted before the laser beam reaches the sample in order to improve the monochromatic character of the wavelength. Interference filters block all but a small number of incoming wavelengths centered at the laser resonance. Holographic notch filters are inserted after the Raman light is scattered from the sample in order to filter the Rayleigh line from the detected light. These filters are especially important in single monochromator detection schemes where the Rayleigh line cannot be removed as it is in double or triple monochromators.

The recent development of the liquid crystal tunable interference filter (LCTF) has provided significant performance advantages over nontunable systems. The LCTF is useful for high-definition Raman imaging because the color of the transmitted light can be electronically controlled from the visible to the near infrared wavelengths of the spectrum. When a LCTF is paired with a CCD detector, the result is a multispectral, high-resolution image.

VI. BASIC EXPERIMENTAL TECHNIQUES

Though normal Raman spectroscopy is a very selective technique for chemical analysis, there are some serious experimental disadvantages related to the sensitivity, large fluorescence interference, and lack of time resolution of the technique. These weaknesses have been addressed in the creation of new Raman-based techniques. The weak Raman signals due to inherently small Raman scattering efficiencies has been addressed by resonance Raman, surface-enhanced Raman and SPP-Raman techniques. Fourier transform-Raman spectroscopy and confocal Raman microscopy address the disadvantage of

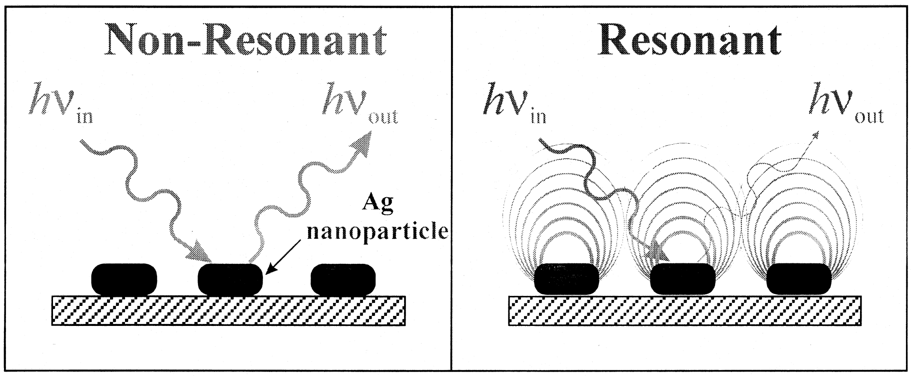


FIGURE 5 Schematic depiction of nonresonant and resonant impinging radiation on a metallic roughness feature. The resonant condition demonstrates the concept of the LSPR.

fluorescence interference. Time-resolved resonance Raman spectroscopy allows kinetic studies of transient reaction intermediates. Nonlinear Raman techniques add yet another dimension to Raman spectroscopy; these techniques are addressed in Section VII.

A. Surface-Enhanced Raman Spectroscopy

Though Raman spectroscopy gives detailed chemical and conformational information, the small scattering intensity has prohibited widespread use. Experiments done in the mid 1970s demonstrated that molecules adsorbed to a roughened metal surface generated anomalously large Raman intensities. This phenomenon is known as surface-enhanced Raman spectroscopy (SERS). SERS enhancements of greater than 10^6 are seen regularly when molecules are adsorbed to roughened silver, gold, and copper surfaces. Enhanced signals have also measured from molecules on lithium, sodium, potassium, cesium, rubidium, aluminum, gallium, indium, platinum, and rhodium.

Today, the SERS community generally agrees that the 10^6 enhancement factor can be attributed to the combination of two separate enhancement mechanisms, the chemical enhancement mechanism (CHEM) and the electromagnetic enhancement mechanism (EM).

The CHEM enhancement mechanism, thought to contribute an enhancement factor of 10^2, asserts that a charge-transfer state is created between the metal and adsorbate molecules. This mechanism is site-specific and analyte dependent. The molecule must be directly adsorbed to the roughened surface in order to experience the CHEM enhancement.

The EM enhancement mechanism contributes greater than 10^4 times enhancement over normal Raman scattering. In order to understand the EM enhancement, one must consider the size, shape, and material of the surface's nanoscale roughness features. If the correct laser wavelength strikes a metallic roughness feature, the plasma of conduction electrons will oscillate collectively. Because this collective oscillation is localized at the surface of this plasma of electrons, it is known as a localized surface plasmon resonance (LSPR). The LSPR allows the resonant wavelength to be absorbed and scattered, creating large electromagnetic fields around the roughness feature. The LSPR is depicted in Fig. 5. If a molecule is placed within the electromagnetic fields, an enhanced Raman signal is measured. Current estimates suggest that these electromagnetic fields reach further than 65 nm from the metal surface.

The discovery and understanding of SERS was important not only because it made Raman a more viable analytical method but also because it introduced the concept of surface-enhanced spectroscopies in general. With the SERS precedent, surface-enhanced resonance Raman spectroscopy (SERRS) and surface-enhanced hyper-Raman spectroscopy (SEHRS) have both been discovered and put to use as analytical tools. In fact, enhancement factors as large as 10^{12} have been measured in SEHRS experiments (see Section VII.B). This immense enhancement was only recently surpassed by the 10^{14} enhancement measured in single molecule SERS (see Section X).

B. Surface-Enhanced Resonance Raman Spectroscopy

Surface-enhanced resonance Raman scattering (SERRS) is a combination of the aforementioned CHEM and EM enhancement mechanisms with fulfillment of a resonance condition. This resonance condition is satisfied when an incoming photon instigates a molecule's electric dipole allowed transition into an excited state, rather than an intermediate state (see Fig. 1). In a standard SERRS experiment, the molecules are first adsorbed to a Ag, Cu, or Au substrate, and then, the absorption spectrum is measured.

A laser wavelength close to the molecule's electronic absorption band is chosen as the Raman excitation wavelength. When the Raman spectra are captured, the signals are 10^9–10^{11} times greater than normal Raman scattering signals. This large enhancement factor is attributed to the cumulative effects of the $>10^6$ SERS enhancement and the 10^3–10^5 resonance Raman enhancement.

C. SPP-Raman

Just as a standing wave is generated when a SERS active surface with nanoscale roughness is irradiated, a propagating wave is generated when a smooth SERS-active surface is irradiated. This traveling wave is known as a Surface Plasmon Polariton (SPP). The SPP can be located by measuring the Attenuated Total Reflectance (ATR) of a thin film surface; the SPP is seen as a deep minimum in the reflectivity versus incident angle dependence when the surface is excited by p-polarized light. SPP measurements give information about the thickness and refractive index of ultrathin films.

Even if two systems are found to be optically identical (i.e., refractive index and film thickness are the same), a sensitive tool like Raman spectroscopy can show how they are different at the molecular level. Although SPP-Raman enhancement (10^3–10^4) is smaller than SERS enhancement, SPP-Raman is necessary when nanoscale roughness would complicate or disturb the adsorbate/substrate system. By coupling the SPP technique with Raman spectroscopy, it is possible to obtain information about the thickness, refractive index, and chemical identity of thin films.

A standard experimental apparatus uses a thin metal (Ag or Au) film on one side of an equilateral triangular, hemicylindrical, or hemispherical prisms. The metal film is placed in contact with the adsorbate molecules to be investigated. Laser light is focused through one side of the prism, to the back of the metal film. Raman scattered light is collected from the front surface of the Ag or Au film and focused unto the monochromator/detector. The total enhancement is dependent on the thickness and dielectric properties of the adsorbate, media, and metal. When the interaction between the SPP and the Raman excitation wavelength is maximized by adjusting either the incident or scattering angle, the Raman signal intensity is maximized.

D. Fourier Transform Raman Spectroscopy

One major limitation to the routine use of Raman spectroscopy lies in the interference caused by the concomitantly excited fluorescence signals. The simplest solution for this problem is to avoid the absorption process giving rise to fluorescence by irradiating the sample with near infrared (NIR) wavelengths rather than visible wavelengths. Until the development of Fourier transform (FT)-Raman in 1986, this solution was not implemented because longer excitation wavelengths translate into smaller Raman peaks intensities by the $1/\lambda_{ex}^4$ dependence of the Raman scattering intensity. For example, using a 1064 nm Nd:YAG laser instead of a 514.5-nm Ar^+ laser produces 18 times less Raman scattering intensity. FT-Raman makes use of the high throughput and wavelength multiplexing features of a Michelson interferometer in order to compensate for the excitation wavelength dependence loss. Current FT-Raman instruments typically use a diode-pumped, air-cooled cw Nd:YAG laser source, a broad-range quartz beamsplitter, gold-coated optics, and an InGaAs or Ge diode detector. The spectral range for a modern commercial FT-Raman instrument is $3600 - 70\ \text{cm}^{-1}$ Stokes shift and -100–$-2000\ \text{cm}^{-1}$ Anti-Stokes shift. FT-Raman advantages include fluorescence elimination and accuracy in frequency measurements.

In the FT-Raman instrument, all frequencies are measured simultaneously so that the entire spectrum can be acquired in a short period of time. The replacement of monochromator slits with large apertures also improves the overall spectral throughput. Most FT-Raman systems utilize an internal He–Ne optical standard to ensure wavelength accuracy. The wavelength accuracy of the data is an order of magnitude higher than that from scanning instrumentation, allowing the kind of sophisticated data manipulation that has been developed by FTIR spectroscopists. Thus, Raman data can be obtained from samples such as industrial polymers and dyes as well as pharmaceutical and petroleum products, which were often inaccessible to earlier generations of dispersive Raman spectroscopy using blue-green visible excitation.

E. Confocal Raman Microscopy

In another application of a multichannel detector, the Raman microscope provides an image of a microscopic sample according to the Raman-active frequencies of the components making up the sample. There are two main classes of multichannel detectors: a diode array, which is a linear array of minuscule diodes each of which functions as a miniature photomultiplier tube, and an image intensifier tube, which is akin to a low-light-level television camera. Diode arrays can collect information in only one dimension, whereas an image intensifier permits collection over the photoactive surface of the intensifier—that is, in two dimensions. Hence, an image intensifier affords an entire two-dimensional picture and gives rise to the potential for performing Raman microscopy. In the Raman

microscope, the laser beam is focused into a sample and the focal point is transferred by a microscope and a monochromator to the active surface of an image intensifier. By adjusting the monochromator, it is possible to select different positions of the sample according to their Raman peaks.

The advantages of Raman microscopy can be further extended by adding a confocal pinhole to create a confocal Raman microscope. By placing a screen with a pinhole on one side of the lens system, only the light from the focal point of the lenses will pass through the pinhole to the detector. Any light resulting from outside the lens' focal point will be out of focus when it reaches the pinhole (see Fig. 6 inset). The focal point of the objective lens and the pinhole are known as conjugate points because the lenses form an image exactly at the pinhole. Because the pinhole is conjugate to the focal point of the lenses, this is known as "confocal" microscopy.

Figure 6 illustrates the confocal Raman microscope. In this apparatus, laser light is reflected by scanning mirrors, through the microscope, to the sample. The scanning mirrors are used to investigate the sample pixel by pixel so that a Raman image can be obtained. The scattered light reflects back through the microscope, and the longer wavelength (scattered) light passes through a dichroic mirror, focuses at the pinhole, and registers on the detector. Femtoliter sample volumes can be investigated with this method.

F. Time-Resolved Resonance Raman Spectroscopy

Time-Resolved Resonance Raman Spectroscopy (TR^3S) is a technique used to get structural, kinetic, and molecular interaction data from chemical and biological systems by recording resonance Raman spectra in a short time span. Using TR^3S, a transient molecular species can be analyzed by (i) monitoring the frequency of vibrational modes to garner structural information, (ii) monitoring peak intensity changes as a function of time to get kinetic information, and (iii) studying the dynamics of the vibrational modes while changing environmental parameters to get information about molecular interactions.

A widely used TR^3S configuration is the pump–probe experiment. A pump pulse initiates a photochemical or photophysical event while the second pulse, a probe pulse, generates resonance Raman scattering. In order to produce a high Raman signal-to-noise ratio, the lasers typically have 10- to 1000-Hz repetition rates with picosecond—100 femtosecond pulse widths. Because these two pulses are independent of one another, both the time delay between the pulses and their excitation wavelengths can be selected. This allows for maximum time resolution as well as optimization of the resonance wavelength. By varying the time delay between the two laser pulses, a reaction's kinetic profile can be acquired by measuring the resonant Raman spectrum for a sequence of delays. Recent TR^3S experiments have explored such issues as the dynamics of ring-opening reactions, the mechanisms of photobiological systems, and electron-transfer in radical reactions.

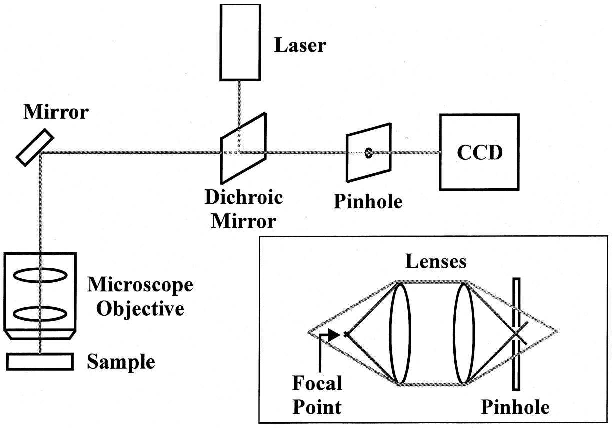


FIGURE 6 Schematic diagram of a confocal Raman microscope.

VII. NONLINEAR RAMAN EFFECTS

In Eq. (1) it was assumed that the induced dipole varied in a linear fashion with the electric field. However, for electric field intensities above 10^9 V/m, as are often produced by pulsed lasers, the linear dependence breaks down. New spectroscopic phenomena arise from the nonlinear interaction of a system with intense monochromatic radiation. Each of the four examples considered here involves changes in wavelength of the radiation as a result of interaction with the system and can be considered to be a variant of the Raman effect. The four examples are the hyper-Raman scattering, the surface-enhanced hyper-Raman scattering, stimulated Raman scattering, and coherent anti-Stokes Raman scattering (CARS).

A. Hyper-Raman Spectroscopy

Hyper-Raman scattering arises from illuminating a system with a focused, pulsed laser that has an irradiance just above the threshold for nonlinear interaction. If the incident radiation is of frequency $1\nu_0$, the scattered radiation is found to include frequencies of the type $2\nu_0$ and $2\nu_0 \pm \nu_1$, where ν_1, is a frequency associated with a transition between two levels of the scattering molecules. The hyper-Raman radiation arises from the second-order nonlinear induced dipole. Thus, Eq. (1) can be expanded to

$$\mu = \alpha \cdot E + \frac{1}{2}\beta \cdot E \cdot E + \cdots, \qquad (24)$$

and the hyper-Raman scattering is controlled by the hyperpolarizability tensor β. The symmetry factors governing the factors of the β tensor are not the same as those for the components of the α tensor. As a consequence, the vibrational selection rules for hyper-Raman scattering are significantly different from those for linear Raman scattering. For example, the torsional vibration in $CH_2{=}CH_2$ is neither IR nor Raman active but is hyper-Raman-active.

B. Surface-Enhanced Hyper-Raman Spectroscopy

Surface-enhanced hyper-Raman scattering (SEHRS) is the analog of hyper-Raman scattering just as SERS is the analog of normal Raman scattering. By adsorbing hyper-Raman-active molecules to a nanoscale roughened Ag, Cu, or Au surface or the corresponding colloids, enhancement factors as large as 10^{12} have been measured. The main utility of the SEHRS technique lies in the new chemical information available as a consequence of the different selection rules. While low-symmetry molecules will have very similar SER and SEHR spectra, high-symmetry molecules will have spectra that show new vibrational features. The combination of three factors: (1) the 10^{12} enhancement factor of SEHRS; (2) solid-state, femtosecond lasers with high repetition rate, high peak power, and low average power; and (3) high-efficiency multichannel detectors portends a new era in vibrational spectroscopy using SEHRS to extract new structural information from important chemical and biological systems.

C. Stimulated Raman Effect

When monochromatic radiation from a pulsed laser of sufficiently large irradiance is incident on a scattering system, hyper-Raman scattering is replaced by a different phenomenon: stimulated Raman scattering. In a typical experimental configuration, the laser beam is focused into the sample and the scattering is observed in the forward direction, along the axis of the laser beam direction, and at a small angle to this direction. The forward-scattered radiation is found to consist of the incident frequency ν_0 and Stokes and anti-Stokes frequencies of the general formula $\nu_0 \pm n\nu_1$, where ν_1 is usually associated with just one Raman-active vibration of the scattering molecules and $n = 1, 2, 3, \ldots$. For example, if liquid benzene is illuminated, the Stokes and anti-Stokes shifts are all exact multiples of 992 cm^{-1}, which represents the strongest feature in the normal Raman spectrum of benzene.

Stimulated Raman scattering differs from normal Raman scattering in its angular dependence, intensity, and frequency distribution. A major fraction of the incident light is converted to radiation at $\nu_0 \pm n\nu_1$. In benzene, 50% of the incident radiation at ν_0 may be converted to Stokes radiation at $\nu_0 - \nu_1$. The high-conversion efficiency of stimulated Raman scattering, taken with the coherent nature of the scattering, gives rise to the possibility that the effect can be used to generate laser-like sources at new frequencies. For example, H_2 or D_2, gas under pressure is used routinely to shift the incident radiation by 4160 or 2942 cm^{-1} (the H_2 or D_2, stretching frequencies, respectively) to gain access to the deep ultraviolet, vacuum ultraviolet, and mid-infrared spectral regions that are not easily accessible in other ways. Stimulated Raman scattering in H_2 and D_2 has been used extensively in ultraviolet resonance Raman spectroscopic studies of proteins.

D. Coherent Anti-Stokes Raman Spectroscopy

The CARS technique involves the use of two tunable dye lasers, a pump laser and a Stokes laser, set at frequencies ν_P and ν_S, respectively. If these two light beams cross in the sample at the phase-matching angle θ, coherent anti-Stokes emission at $\nu_{AS} = 2\nu_P - \nu_S$ is generated through the third-order linear polarization. In fact, $I_{CARS} \sim I_P^2 I_S$. The laserlike beam due to ν_{AS} is greatly enhanced when the frequency interval, $\nu_P - \nu_S = \Delta$, is equal to a Raman-active molecular vibrational frequency. Thus, vibrational Raman spectra are obtained by fixing the frequency of ν_P and varying the frequency of ν_S. Fluorescence rejection occurs in a CARS experiment because the signal beam, ν_{AS}, is spatially and temporally removed from the fluorescence signal.

In previous years, the major application of CARS was in the analysis of gases and gas mixtures, but recent advances have demonstrated that CARS can be applied to nongaseous systems as well. A problem in the application of CARS to molecules in solution is interference from the solvent, which can contribute a background emission resulting from the third-order susceptibility of the solvent. To a certain extent, this problem may be overcome when the CARS experiment is carried out under resonance conditions. As the ν_P beam frequency approaches an electronic transition of the solute, the solute's CARS signal is resonance enhanced whereas the background emission remains unchanged.

VIII. CHEMICAL AND BIOCHEMICAL APPLICATIONS

A. Chemical Applications

Five representative applications in which Raman spectroscopy has proved to be very powerful are cross-sectional imaging of live cells with CARS, investigation

of industrial polymers with Fourier Transform (FT) Raman Spectroscopy, the study of self-assembled monolayers (SAMs) with SERS, the assignment of vibrational bands by the combination of SERS and SEHRS, and SPP-enhanced Raman studies of carbon clusters.

1. CARS Cross-Sectional Imaging

Recent CARS experiments have targeted cross-sectional imaging by making use of solid state femtosecond lasers with λ_p and λ_s in the NIR region. The long wavelength pulses give two major advantages over visible excitation: (i) the excitation wavelengths are far from electronic transition bands, so the background signals are small, and (ii) Rayleigh scattering is minimized in heterogeneous samples, allowing deeper penetration into the sample volume. By collinearly coupling the two pulses into an optical microscope and through a high numerical aperture objective, the excitation pulses can be focused tightly onto the sample. Because CARS is a nonlinear optical process, only a small volume of the sample is excited. The advantages of small volume excitation include background signal rejection, decreased photodecomposition of the sample, and the ability to section a three-dimensional object by changing the focal plane of the excitation pulses. The consequence of a long wavelength, small volume excitation is that three-dimensional imaging of live cells has been achieved. Figure 7 shows CARS images of live, unstained bacteria tuned to the Raman shift of 2878 cm^{-1}, an aliphatic C–H stretching band. The lipid bilayer of the bacterial cell membrane is rich in aliphatic C–H bonds. CARS will find more application in coming years because it gives vibrational contrast and high sensitivity, but the power levels are tolerable to living cells.

2. FT-Raman Analysis of Polymers

The spectroscopic advances made in the field of polymer chemistry demonstrate the impact of the FT-Raman technique. Polymer samples are historically plagued by large fluorescence backgrounds and sample degradation under the heating effects of visible wavelength laser irradiation. In fact, estimates suggest that up to 95% of polymer samples cannot be examined by normal Raman spectroscopy. Until the advent of FT-Raman, most polymer samples were investigated by infrared (IR) spectroscopy. Often, the strenuous sample preparation required for IR investigations alters the polymer structure, thus sacrificing an accurate spectroscopic view. The high symmetry of most polymer samples also acts as a disadvantage in IR experiments because many symmetric stretching mode vibrations are IR inactive. Often, IR spectra are missing all information regarding the homonuclear polymer backbone.

FT-Raman addresses many of the disadvantages of IR polymer analysis. First, because sample preparation is minimal for FT-Raman analysis, no valuable information is lost. Second, because the Raman selection rule is based on changing bond polarizability rather than on the changing dipole moment (as in IR spectroscopy), the homonuclear backbone stretches can be seen in Raman spectra. Third, detailed information about the chemical composition, structure, and stereoregularity can be obtained by FT-Raman measurements. Some of the most exciting data gained in FT-Raman investigations of polymers lie in the details of chain conformation due to intermolecular interactions. The chain-packing details help to explain physical characteristics of polymers such as absorbance spectra and transition temperatures.

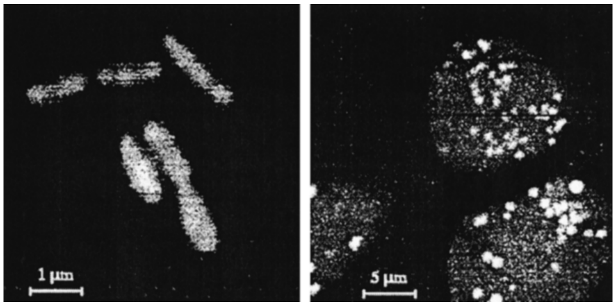


FIGURE 7 CARS image of a live bacterial cell. Imaging was tuned to the Raman shift of 2878-cm^{-1} band (aliphatic C–H stretch). [Reproduced from Zumbusch *et al.* (1999). *Phys. Rev. Lett.* **82**(20), 4142, by permission.]

3. SERS of Self-Assembled Monolayers at the Solid/Liquid Interface

SERS experiments give scientists a window into the complex processes occurring at solid/liquid, solid/gas, solid/UHV, and solid/solid interfaces. When molecules adsorb to a SERS-active substrate, the measured spectra give valuable information concerning the structure and reactivity of the adsorbate molecules. Comparison of SERS spectra to bulk Raman spectra demonstrates any chemical or conformational variance between the free and adsorbed states. This information is important when using SERS to understand complex phenomena such as heterogeneous catalysis and electrochemistry at the molecular level.

As an example, consider the pervasive use of self-assembled monolayers (SAMs) for the surface modification of gold and silver substrates. A recent SERS/SERRS study of the Au film/SAM adsorbate interface demonstrates that thin films of Au(111) on mica substrates, previously thought to be SERS-inactive, is SERS-active without any additional treatment. This discovery is quite

significant because it allows SERS experiments to be executed on undisturbed, unroughened Au(111)/SAM adsorbate surfaces. The impact of better understanding SAM adsorption to a Au(111) substrate is significant because this system acts as a model for other substrate/adsorbate systems.

In the aforementioned experiment, the SERS activity of rough Au films is compared with that of single crystal Au(111) films. Figures 8A and B show nanometer-scale AFM images of a rough Au surface and a Au(111) surface before SAM modification. The corresponding SERRS of SAMs on these surfaces are in Figs. 8C and D, respectively. The resonance Raman condition is a result of overlap between the adsorbate molecule's electronic absorption band and the excitation wavelength. Although not shown within this text, SERS spectra were also measured for each sample. With these two spectra, the surface resonance Raman (SRR) enhancement ($\sim 10^2$) could be separated from the EM enhancement ($\sim 10^3$). This weak EM enhancement, compared to the standard EM contribution of 10^4–10^5, is attributed to the $\sim$100 nm wide, atomically flat islands that comprised the surface of Au(111)/mica samples shown in Fig. 8B or the roughness features between the terraces of the Au(111) surface. Future experiments will address these hypotheses.

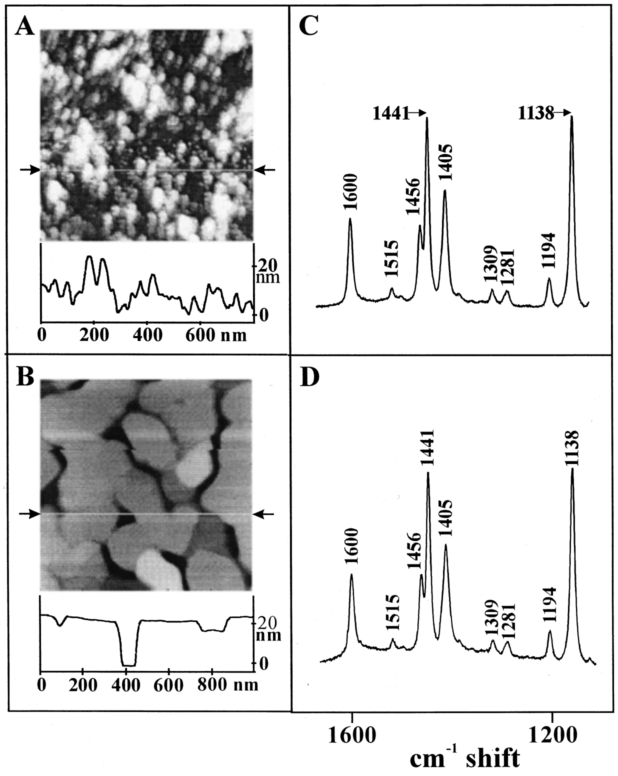


FIGURE 8 SERS-active surface topography and SERRS spectra for SAM on rough Au and Au(111) substrates. (A) depicts the AFM image and line scan of the rough Au substrate. (B) depicts the AFM image and line scan of the Au(111) substrate. (C) is the SERRS spectrum of SAM adsorbed to the surface shown in (A). (D) is the SERRS spectrum of SAM adosrbed to the surface shown in (B). [Reproduced from Caldwell *et al.* (1994). *Langmuir* **10,** 4109, by permission.]

4. SEHRS Assisted Assignment of Vibrational Bands

A particularly salient example of the complementarity of normal Raman spectroscopy with other techniques is expressed in a recent publication of the combined data from infrared, normal Raman, SERS, SEHRS, and theoretical predictions for one molecule. First, *ab initio* theoretical predictions were made for the vibrational characteristics of *trans*-1,2-bis(4-pyridyl)ethylene (BPE) at the Hartree-Fock 6-31G* level. When the spectra were collected, comparisons were made between the theoretical and experimental results as well as among the different spectra. Based on the known selection rules for each spectroscopy and the matching of wavenumber shifts to theoretical predictions, all vibrational bands were assigned.

Though this study only gives exact information for BPE, it demonstrates the power of combined theoretical/spectroscopic characterization. The development, in recent years, of *ab initio* electronic structure calculations of high accuracy has revolutionized theoretical chemistry. The ability to efficiently calculate vibrational properties from first principles has or will completely replace normal coordinate analysis.

Figure 9 shows the SER and SEHR BPE spectra as well as the theoretical prediction for each. Although the hyper-Raman scattering efficiency is eight orders of magnitude smaller than the linear Raman scattering efficiency, the SEHR signals are only three orders of magnitude smaller than the SERS signals. This translates to 10^{12}-fold enhancement over normal hyper-Raman scattering. There has been renewed interest in hyper-Raman spectroscopy due to this immense SEHR enhancement. Vibrational modes that are active in only hyper Raman spectroscopy can be measured reliably for the first time.

5. SPP-Enhanced Raman Spectroscopy of Carbon Clusters

SERS is generated from molecules adsorbed to a roughened metal surface, but not all molecules can maintain their structure and function in direct contact with a metal surface. Both biomolecules and atomic clusters fall into this category. One solution to this limitation is to put a spacer molecule (such as the SAM suggested in the previous section) between the metal surface and the molecule

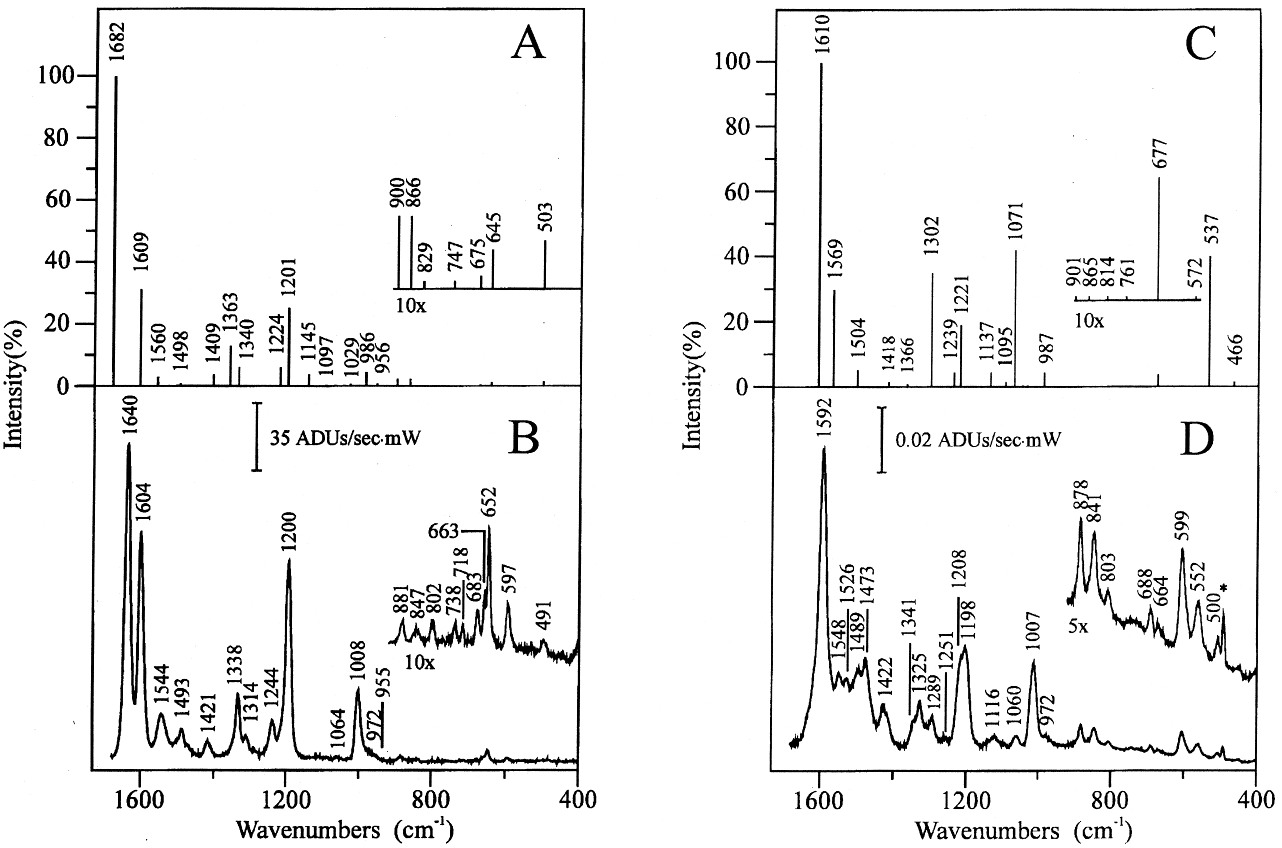


FIGURE 9 Theoretical SER (A) and SEHR (C) and experimental SER (B) and SEHR (D) spectra of BPE on a roughened silver surface. [Reproduced from Yang *et al.* (1996). *J. Chem. Phys.* **104**(11), 4313, by permission.]

of interest. In cluster science, a common solution to the problem of high reactivity is to embed the clusters inside solid matrix of condensed inert gas. SPP-enhanced Raman spectroscopy allows Raman spectra to be measured from smooth surfaces.

A recent SPP-resonance Raman experiment explored the structural conformations of several carbon clusters. The C_{16}, C_{18}, and C_{20} clusters were created by laser ablation of a graphite rod and then deposited into a N_2 matrix on the silvered SPP prism surface. After finding the SPP resonance condition (as described in Section VI.B), Raman spectra were collected with six different excitation wavelengths. SPP-Raman enhancement is operative in all six spectra. The strong dependence of the Raman spectra for the C_{20} cluster on λ_{ex} shown in Fig. 10 indicates that enhancement due to RRS is simultaneously operative. It is important to emphasize that these signals shown in this figure are obtained from ca. 10^{10} clusters in the laser focal spot. Upon comparing the Raman peak frequencies for all three carbon clusters to theoretical predictions, the researchers were able to hypothesize that all three carbon clusters adopt either a linear chain or polyacetylene ring conformation, but not a fullerene or bowl structure.

B. Biochemical Applications

A major advantage of Raman spectroscopy for the analysis of biomolecules stems from the fact that water has a weak Raman spectrum. Spectra can be recorded for aqueous solutes at 10^{-1}–10^{-2} M with little interference from the solvent. For a chromophore under the RR condition the accessible concentration range becomes 10^{-4}–10^{-6} M. Moreover, the intensity enhancement associated with the RR effect confers the important advantage of selectivity, allowing one to observe selectively the vibrational spectrum of a chromophore that is just one component of an extremely complex biological system. Because many biomolecules have chromophores with an ultraviolet (UV) resonance condition, one may also selectively excite a chromophore by irradiating these molecule with UV light. This technique is known as Ultraviolet Resonance Raman Spectroscopy (UVRRS). In recent years, Raman difference spectroscopy (RDS) has been developed in

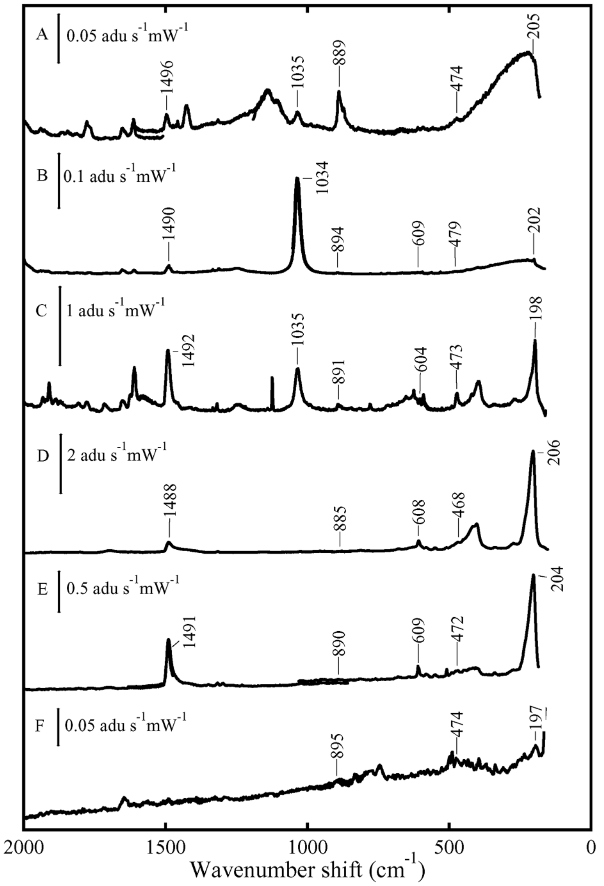


FIGURE 10 SPP-Raman spectra of matrix-isolated C_{20} clusters over the energy region 100–2000 cm^{-1} for excitation wavelengths of (A) 457.9, (B) 488, (C) 514.5, (D) 635, (E) 647, and (F) 670 nm. [Reproduced from Ott, *et al.* (1998). *J. Chem. Phys.* **109**(22), 9653, by permission.]

order to deconvolute the complicated spectra of biological macromolecules.

Three major classes of biomacromolecules have been studied by normal (nonresonance) Raman spectroscopy: proteins, nucleic acids, and lipids and membranes. The type of information obtained for each class can be summarized as follows.

1. *Proteins*. Quantitation of polypeptide conformation, α-helix, β-sheet, β-turns, etc.; characterization of cysteine–SH side chains; conformation of disulfide –S–S– linkages; strength of hydrogen bonds to tyrosine–OH; exposure to hydrophobic/hydrophilic environments of tryptophan side chains
2. *DNA and RNA*. Quantitation of the (deoxy)ribose-phosphate backbone conformation and base composition; observation of base pairing and base stacking (often different classes of bases can be monitored separately); H–D exchange in bases
3. *Lipids and membranes*. Interchain interactions and melting behavior of the lipid aliphatic chains; effects of chemical perturbants (e.g., cholesterol) and proteins on melting behavior; lipid head-group conformation

RR spectra have been obtained from chromophores in live tissue and from bacterial cells under physiological conditions. However, the bulk of the studies in the past 20 years has been on purified materials: the type of information gained is summarized here.

1. *Heme proteins*. Oxidation and spin state of the heme iron using porphyrin marker bands; detailed chemistry and hemoglobin–CO photolysis on the nano- and picosecond timescales; quaternary structures; chemistry of Fe–ligand bonds
2. *Visual pigments and bacteriorhodopsin*. Retinal conformation in the photocycle intermediates; state of protonation of the retinal-protein Schiff base
3. *Metalloproteins*. Chemistry of the ligands around the chromophoric metal center

Most of the other naturally occurring biological chromophores, such as flavins and carotenoids, have now been studied. UVRRS is a technique that complements visible and infrared Raman spectroscopies by providing sensitive, selective information about a small number of vibrational bands. In order to make maximal use of the UV Raman technique, one must carefully choose the species in solution so that the Raman cross sections and bands of the solution do not compete with those of the analyte being studied. With a well-chosen system, this technique can give information about the structure and dynamics of small molecules and functionalities and chromophores on larger molecules. UV Raman is most useful because many molecules have absorption bands deep in the UV region, and fluorescent interference is greatly decreased because fluorescence does not usually occur in species with an excitation band below 260 nm.

In classic UV Raman systems, an excimer laser was used to pump a frequency-doubled dye laser which was then directed at the liquid sample as it was jetted along a guided path. Recently, UV Raman spectroscopy has become more feasible due to the advent of a frequency-doubled Ar^+ laser that has five excitation wavelengths with substantial power in the UV region. This laser yields better signal-to-noise ratios and allows for UV Raman investigation of solid samples. The most recent UV Raman studies examine protein–nucleic acid interactions, protein–protein interactions, and the structure and behavior of proteins, DNA, and neurotransmitters in aqueous solutions.

However, many biological sites of importance do not contain a suitable chromophore, and for these, the RR labeling technique was developed. An RR label, usually mimicking a natural component, is introduced into the system as a reporter group, and there has been success in using chromophoric ligands to study antibody-haptcn, enzyme-inhibitor (and drug), DNA–drug, and cell–dye interactions. Moreover, the labeling technique can aid the understanding of the molecular details of enzymolysis. The hydrolysis of thionoesters (of the type $RC(=O)NHCH_2C(=S)OCH_3$) by the enzyme papain occurs via the formation of a transient dithioester $RC(=O)NHCH_2C(=S)$–S–papain (the thiol sulfur belongs to a cysteine side chain in the active site). The dithioester absorbs at 315 nm, and thus, the label is generated at the time and location of catalysis. The 324-nm excited RR spectrum in Fig. 11 shows how different spectral features monitor conformations in different parts of the enzyme–substrate complex. By this means it is possible to monitor the critical events in the bonds undergoing catalytic transformation in a complex of molecular weight 24,000.

Raman analysis of large macromolecules, such as proteins, often yields complicated, spectrally crowded results.

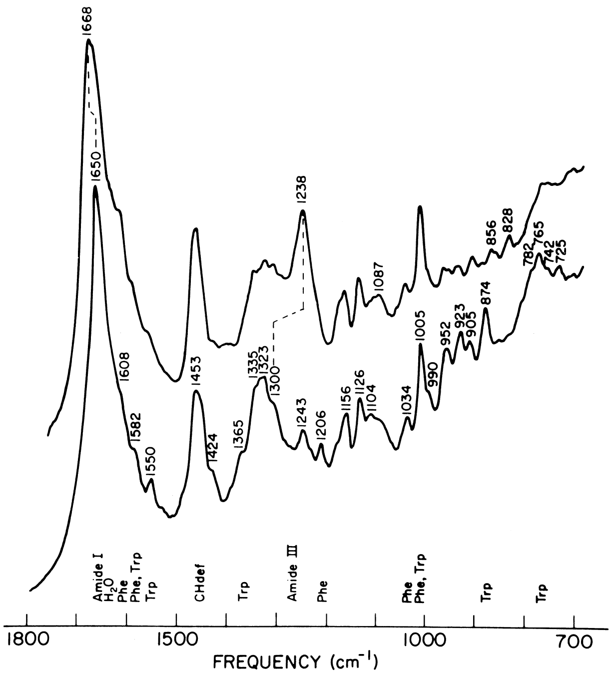


FIGURE 11 Monitoring the group (and its neighbors) undergoing transformation in an enzyme's active site. The RR spectrum shown is of the enzyme–substrate transient $PhC(=O)NHCH_2(=S)S$-papain; 324-nm excitation. [Reproduced from Carey, P. R. (1992). "Raman Spectroscopy," *In* "Encyclopedia of Physical Science and Technology," 2nd ed., Academic Press, New York, by permission.]

Resonance Raman labeling is used in order to highlight the spectrum of small portions of the large molecule. This technique has limited applications to protein molecules because the chromophore dominates the observed spectrum and not all proteins will accept chromophore addition. A more generally applicable method known as RDS has become technically feasible in recent years to study events such as protein–ligand binding, enzymatic catalysis, and protein assembly.

In RDS, spectra of both the normal and the modified version of the macromolecule are collected. By subtracting the normal spectrum from the tagged molecules' spectrum, information about the tagged portions of the molecule is obtained. Two methods used in the tagging process are ligand binding and isotopic editing. By subtracting a bare protein's spectrum from the spectrum of the protein bound to a ligand, the RDS spectrum of the bound ligand is obtained. In isotopic editing, the substitution of an atom in the bond of interest with another stable isotope will cause shifts in the Raman bands for motions involving the edited nucleus. By subtracting the normal spectrum from the isotopically edited spectrum, Raman peaks will be seen only for the modes of vibration involving the isotopically edited atoms.

IX. THE SOLID STATE

Raman spectroscopy of the solid state differs from that of gases or liquids. A fluid is usually considered to be an assembly of noninteracting, randomly oriented molecules. In contrast, in a solid, i.e., a molecular crystal, the molecules have essentially fixed orientation with respect to the crystal axis. Thus, the molecules lose the rotational and translational degrees of freedom found in the free molecule. In the molecular crystal, these degrees of freedom are replaced by so-called external vibrations: torsional motions of the molecule about its axis at the lattice site (librations) and restricted translational excursions within the lattice site. The external vibrations give rise to many new features in the low-frequency region of the spectrum. In addition, the vibrational bands seen for the free molecule (the internal vibrations) can be perturbed in the crystal. These vibrational bands may be split, for example, due to the symmetry of the crystal site, interactions with other molecules in the unit cell, or interaction between the vibrations of the free molecule and the external vibrations in the crystal.

The potential for splitting the internal vibrations and the appearance of features due to external vibrations means that the Raman spectrum of the molecular crystal may be quite different from that of the free molecule. The degree of difference depends on the symmetry and crystal sites of the molecule and on the strength of the coupling between the motions in the crystal. If the intermolecular coupling is

much weaker than the intramolecular forces, the spectrum of the crystal may be only slightly perturbed from that of the free molecule. However, this is often not the case; examples of strong intermolecular coupling are found in molecular crystals. At the extreme end of the scale, the analysis of the Raman spectra of ionic crystals such as NaCl or KBr or of covalent crystals such as diamond, depends on a radically different approach in which the motions of the crystal lattice as a whole are considered. Such a treatment is beyond the scope of this article. A few examples will be given to emphasize the range and value of Raman spectroscopy in solid-state research.

A. Symmetry of Vibrations from Single-Crystal Spectra

Raman spectroscopy offers a valuable means of unambiguously determining the symmetry of vibrations for single oriented crystals. Planned selection of the direction and polarization state of both the incident light and the observed Raman scattering enables the magnitude of each tensor component [Eq. (2)] to be determined. By this means, the symmetry classes of the Raman-active vibrations can be identified.

A convenient notation for describing polarization data from single crystals is due to S. Porto. This involves four symbols (usually chosen from x, y, z axis designations), which define the propagation direction of the incident radiation, the direction of the $\boldsymbol{E}$ of the incident radiation, the direction of the $\boldsymbol{E}$ of the scattered radiation being examined, and the direction of propagation of the scattered radiation. The second and third symbols are placed in parentheses, for example, $z(xy)x$, and define the components of the scattering tensor.

Crystalline calcite furnishes an instructive example of this notation. Calcite (a form of calcium carbonate) has the point group symmetry D_{3d} and symmetry arguments determine that the vibrational Raman modes will be made up of one vibration of a_{1g} symmetry and four doubly degenerate pairs of vibrations of e_g symmetry. The polarized Raman spectra of calcite are shown in Fig. 12 for four different orientations. For the illumination observation configuration $y(zz)x$, the scattering is proportional to α_{zz}^2. Since this is nonzero only for the a_{1g} mode, only this vibration will be seen in the $y(zz)x$ configuration. Thus, from Fig. 12, the 1088-cm^{-1} feature is immediately identified as the a_{1g} mode.

B. Semiconductors

Raman spectroscopy has been used to obtain valuable information on the properties of semiconductors. There are two major classes of semiconducting materials: intrinsic semiconductors are very high purity (i.e., single elements) and extrinsic semiconductors are doped materials. Raman spectroscopy is an efficient tool used to investigate semiconductor surfaces or interfaces made up of a semiconductor and a metal or an insulator, or to characterize the strains brought about by heavy doping in semiconductors. In one instance, Raman scattering from an extrinsic semiconductor can give rise to a tunable laser source. The example system involves a wide-band-gap II–VI semiconductor, such as InSb, subjected to a strong magnetic field at low temperature. Under these conditions, interaction with a photon causes two electrons in the semiconductor to change their energy levels. The overall result is the change of electron spin quantum numbers ΔS of ± 1. That is, one spin flip. The spin–flip energy is proportional to the strength of the magnetic field and is approximately 7 cm^{-1} when the magnetic field is $8T$. Thus, intense peaks are observed at 7 cm^{-1} in the Stokes and anti-Stokes Raman spectra. The dependence of the peak position or magnetic field offers the possibility of constructing a tunable, spin–flip laser. For example, the 943-cm^{-1} output of a pulsed CO_2, laser may be tuned from 700 to 1100 cm^{-1} by using the magnetic field dependent spin–flip of InSb.

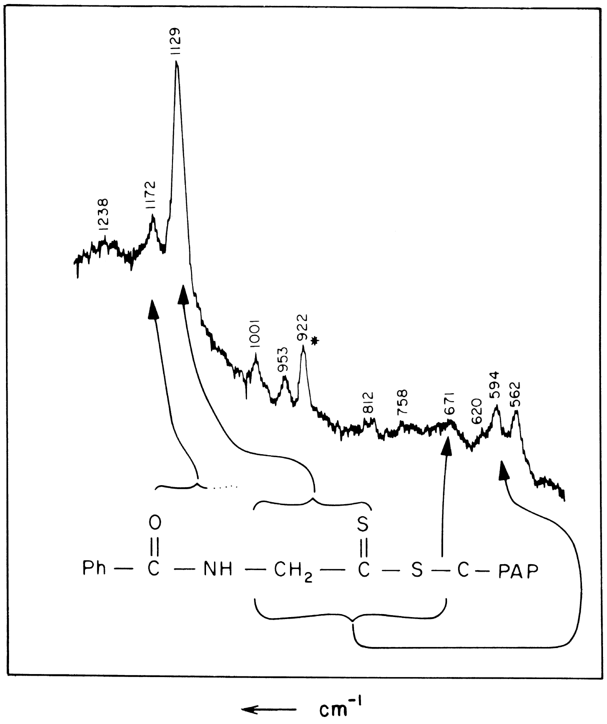


FIGURE 12 Polarized Raman spectra of calcite; 632.8-nm excitation. [Reproduced from Porto *et al.* (1966). *Phys. Rev.* **147,** 608.]

C. Phase Transitions

Structural detail associated with phase transitions can be obtained by recording the Raman spectra of oriented

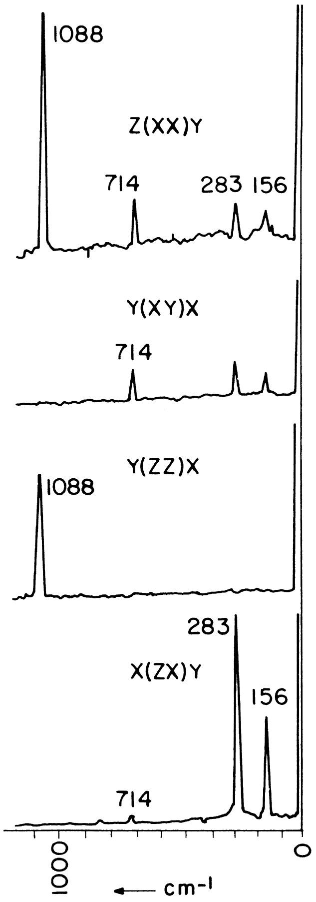


FIGURE 13 Low-frequency Raman spectrum of $NaNO_3$ as a function of temperature. The transition temperature T_c is 548 K. [Reproduced from Carey, P. R. (1992). "Raman Spectroscopy," *In* "Encyclopedia of Physical Science and Technology," 2nd ed., Academic Press, New York, by permission.]

crystals as a function of temperature and pressure. The low wave number spectrum of $NaNO_3$ is shown in Fig. 13. Sodium nitrate undergoes a phase transition at 548 K and, as Fig. 13 shows, this transition is accompanied by disappearance of the band near 20 cm^{-1}. The mode giving rise to this feature is totally symmetric and its appearance at low temperature indicates that the NO_3^- ion is pyramidal, that is, nonplanar. The disappearance of the mode above the phase transition is taken as evidence that the NO_3^- group rotates freely about the axis of symmetry through the N atom, whereas at low temperature, the NO_3^- moiety performs only modest librations about this axis. The band shape changes seen in Fig. 13 show that the amplitude of libration increases with increasing temperature until T_c, the transition temperature. At this point, the NO_3^- is able to cross the potential barrier and the group becomes a free rotor.

X. RECENT DEVELOPMENTS—SINGLE MOLECULE SERS (SMSERS)

Single molecule detection (SMD) is important because it allows for examination of fundamental molecular properties without ensemble averaging. With SMD, it is possible to (i) compare subpopulations of a heterogenous system, (ii) study fluctuating systems in equilibrium, and (iii) examine the reaction pathways and intermediates of individual molecules in systems that are difficult to synchronize. As progressing technology has allowed for creation of increasingly sensitive detectors and the investigation of SERS enhancement mechanisms has advanced, the goal of single molecule surface-enhanced Raman spectroscopy (SMSERS) has been achieved.

While single molecule fluorescence studies have been possible for approximately 10 years, it is only within the last 4 years that SMSERS has arrived. Although Raman cross sections are nine to 14 orders of magnitude smaller than fluorescence cross sections, there are many advantages to using Raman detection for single molecule studies. Fluorescence spectra are hindered by problematic photobleaching, long electronic relaxation times, and little structural information in the acquired data. Raman, on the other hand, avoids photodecomposition because the excitation energy is not in resonance with molecular transitions. The short vibrational relaxation times characteristic of Raman spectroscopy yield 10^3 more photons per unit time than fluorescence; this difference allows for shorter integration times and higher rates for single molecule counting. Raman is also more sensitive than Raman, giving detailed structural and conformational information.

In 1984, Hildebrandt and Stockburger took spectra of as few as 10^4 Rhodamine 6G (R6G) molecules while studying the SERS enhancement mechanism in activated versus unactivated systems. In 1988, Van Duyne *et al.* measured the spectra of 600 pyridine molecules. The Kneipp group detected a signal from 60 R6G molecules in a silver colloid solution in 1994. In this experiment, they also calculated the SERS cross section based on the vibrational state lifetimes, and in order to make those cross sections consistent with the signal level, they calculated that only 0.01% of the molecules were contributing to the SERS signal. After this result was published, theoreticians began hypothesizing different SMSERS enhancement mechanism models such as the existence of "hot spots" on small metal particles and "hot zones" in large fractal structures. Finally, in 1997, the Kneipp group detected a single-crystal violet molecule in a solution of silver colloid clusters. The Nie group also reported single molecule detection in 1997, but this group measured the R6G spectrum on non-aggregated nanoparticles. Figure 14 shows the SMSERS spectra as well as

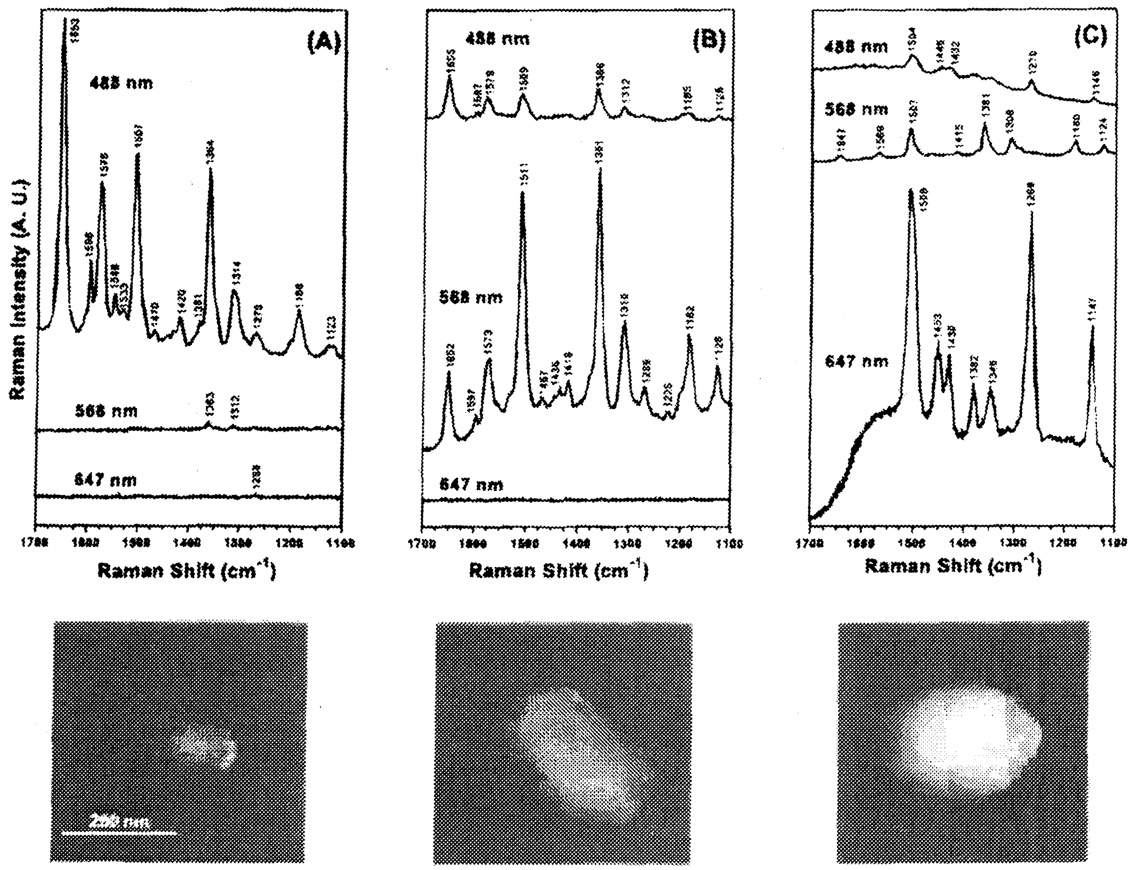


FIGURE 14 Correlated SERS and tapping-mode AFM images obtained from spatially isolated, single Ag nanoparticles. These particles were selected by wide-field screening for maximum enhancement at (A) 488 nm, (B) 567 nm, and (C) 647 nm, respectively. [Reproduced from Nie *et al.* (1998). *JACS* **120,** 8009, by permission.]

atomic force microscope (AFM) images of the SMSERS-active particles. These spectra indicate enhancement factors of up to 10^{14}.

There is a great controversy in the SERS community right now regarding the mechanisms by which this enhancement occurs. Although Kneipp *et al.* (1999) claim that clusters are necessary and Nie and *et al.* (1998) claim that single nanoparticles are necessary, thus supporting different enhancement mechanism models, they are both working in the same nanoparticle size regime. It is exciting that both groups, as well as a few other groups around the world, have achieved single molecule detection using SERS enhancement mechanisms, but there is still a lot of work to be done in order to fully understand the large enhancement factors and the enhancement mechanisms of SMSERS.

SEE ALSO THE FOLLOWING ARTICLES

Atomic Spectrometry • Infrared Spectroscopy • Microscopy • Scattering and Recoiling Spectrometry

BIBLIOGRAPHY

Caldwell, W. B., Chen, K., Herr, B. R., Mirkin, C. A., Hulteen, J. C., and Van Duyne, R. P. (1994). "Self-assembled monolayers of ferrocenylazobenzenes on Au(111)/mica films: Surface-enhanced Raman Scattering (SERS) Response vs. Surface morphology," *Langmuir* **10,** 4109–4115.

Campion, A., and Kambhampati, P. (1998). Surface-enhanced Raman scattering. *Chem. Soc. Rev.* **27,** 241–250.

Chase, D. B., and Rabolt, J. F. (eds.) (1994). "Fourier Transform Raman Spectroscopy," Academic Press, New York.

Hildebrandt, P., and Stockburger, M. (1984). Surface-enhanced resonance Raman spectroscopy of rhodamine 6G adsorbed on colloidal silver. *J. Phys. Chem.* **88,** 5935–5944.

Kneipp, K., Kneipp, H., Itzkan, I., Dasari, R. R., and Feld, M. S. (1999). "Ultrasensitive chemical analysis by Raman spectroscopy," *Chem. Rev.* **99,** 2957–2976.

Laserna, J. J. (ed.) (1996). "Modern Techniques in Raman Spectroscopy," Wiley, New York.

Myers, A. B., and Rizzo, T. R. (eds.) (1995). "Laser Techniques in Chemistry," Wiley, New York.

Nie, S., and Emore, S. R. (1997). "Probing single molecules and single nanoparticles by surface-enhanced Raman scattering," *Science* **275,** 1102–1106.

Ott, A. K., Rechtsteiner, G. A., Felix, C., Hampe, O., Jarrold, M. F., and Van Duyne, R. P. (1998). "Raman spectra and calculated vibrational frequencies of size-selected C_{16}, C_{18}, and C_{20} clusters," *J. Chem. Physics* **109,** 9652–9655.

Schrader, B. (ed.) (1995). "Infrared and Raman Spectroscopy," VCH Publishers, Inc., New York.

Turrell, G., and Corset, J. (eds.) (1996). "Raman Microscopy, Developments and Applications," Academic Press, New York.

Yang, W. H., Hulteen, J. C., Schatz, G. C., and Van Duyne, R. P. (1996). "A surface-enhanced hyper-Raman and surface-enhanced Raman scattering study of *trans*-1,2-bis(4-pyridyl)ethylene adsorbed onto silver film over nanosphere electrodes. Vibrational Assignments: Experiment and theory," *J. Chem. Physics* **104,** 4313–4323.

Zumbusch, A., Holtom, G. R., and Xie, X. S. (1999). "Three-dimensional vibrational imaging by coherent anti-Stokes Raman scattering," *Physical Rev. Lett.* **82,** 4142–4145.

Scanning Probe Microscopy

C. Daniel Frisbie
University of Minnesota

I. Basic Operating Principles
II. Applications of Scanning Tunneling Microscopy (STM)
III. Applications of Atomic Force Microscopy (AFM)
IV. Other SPM Techniques and Applications
V. Future Trends

GLOSSARY

Cantilever A projecting beam supported only at one end, the deflection of which is directly proportional to the force applied normal (perpendicular) to the beam axis.

Constant-current feedback control Regulation of the piezoelectric scanner position based on the magnitude of the electron tunneling current in a scanning tunneling microscope.

Constant-force feedback control Regulation of the piezoelectric scanner position based on the magnitude of the force between the cantilevered probe and sample in an atomic force microscope.

Electron tunneling A quantum mechanical process in which electrons penetrate potential barriers such as a vacuum gap between two metals.

Feedback control Self-corrective optimization of device performance by modification of input based on analysis of output.

Piezoelectric scanner An electrical-to-mechanical transducer that allows precise (angstrom level) positioning and movement in all three dimensions.

Raster scanning A scanning pattern in which a probe is swept side to side in lines over an area from top to bottom.

SCANNING PROBE MICROSCOPY (SPM) refers to a general class of microscopy techniques that produce high-resolution, high-magnification images by scanning a sharp probe back and forth over a specimen. Scanning tunneling microscopy (STM), the first SPM method to be developed, uses a current-sensing probe to record the topography of conducting samples. The spatial resolution of STM topographic images is extremely high, better than 1 Å (0.1 nm or 10^{-10} m), enabling direct visualization of atoms and molecules. Another scanning probe technique, known as atomic force microscopy (AFM), uses a force-sensing probe to track sample topography. The topographic resolution of AFM ranges from 1 to 20 Å depending on the operation mode and the specific sample characteristics; under favorable conditions true atomic resolution can be obtained. A practical difference between the two techniques is that AFM does not require that the sample be electrically conducting.

Since their invention in the 1980s, STM and AFM have provided scientists with exquisitely detailed snapshots of a wide variety of structural and chemical phenomena occurring on surfaces. The impact of STM on our understanding of surface chemistry and physics was apparent immediately. In 1986, Gerd Binnig and Heinrich Rohrer were awarded the Nobel Prize for their invention of the STM only 5 years earlier. The success of STM and AFM has led to the proliferation of scanning probe technology in all aspects of science and industry. New scanning probe microscopies have been developed that allow high-resolution imaging of a large spectrum of sample characteristics in addition to topography, including frictional properties, local electric and magnetic domains, thermal behavior, optical absorption, fluorescence, and birefringence. SPMs are now valued not just as imaging instruments, but as important analysis and metrology tools for examining a range of material properties on nanometer length scales. Scanning probe techniques continue to be developed and to have an impact on virtually every area of technology ranging from molecular biology to semiconductors, surface chemistry, catalysis, and nanofabrication.

I. BASIC OPERATING PRINCIPLES

All scanning probe microscopes have (1) a mechanism for bringing a sharp probe into close proximity with the sample surface, (2) a mechanism for scanning the probe in a raster pattern, and (3) a *feedback mechanism* for regulating the distance between the sample and probe while scanning. These functions, as well as image acquisition and display, are controlled by a computer and associated electronics.

The initial approach of the probe to the sample is accomplished with a simple one-dimensional translator such as a finely threaded screw or piezoelectric element. This translator must be able to move the probe a distance on the order of millimeters toward the sample. Once the probe tip is in contact or near contact with the sample, the probe (or the sample) is scanned in a raster pattern, similar to the way an electron beam is swept back and forth across the screen in a television set. Scanning is accomplished with piezoelectric elements capable of moving the probe in the X and Y dimensions with angstrom-level precision. Tube-shaped piezoelectric scanners are often used in which the probe is attached to one end of the tube. Voltages applied to the tube cause it to bend back and forth, generating the scanning pattern.

In addition to the horizontal (or X–Y) scanning, the tube scanner moves the tip up and down in the Z direction in order to track the surface topography. Surface tracking requires the use of *feedback control* to regulate the probe–sample distance. Different SPM methods are distinguished by the specific feedback mechanism employed and the type of data, in addition to topography, that are recorded. The different feedback mechanisms are best understood by considering specific SPM techniques.

A. Scanning Tunneling Microscopy (STM)

Figure 1A shows a schematic diagram of an STM. The STM probe is a thin metal wire that has been cut or chemically etched to a sharp point. The probe is attached to a piezoelectric tube scanner described above allowing precise sweeping of the probe tip back and forth across the sample. When the probe is within a few angstroms of the substrate, application of a voltage produces a quantum mechanical tunneling current that is extremely sensitive to the probe–sample separation. The tunneling current increases exponentially with decreasing probe–sample distance, and thus it can be used to regulate the probe–sample separation in the Z direction. While the probe is being scanned, the tunneling current at a given applied bias is detected by the feedback control electronics and compared to a “set-point” current that the user chooses (e.g., 1 nA). If the

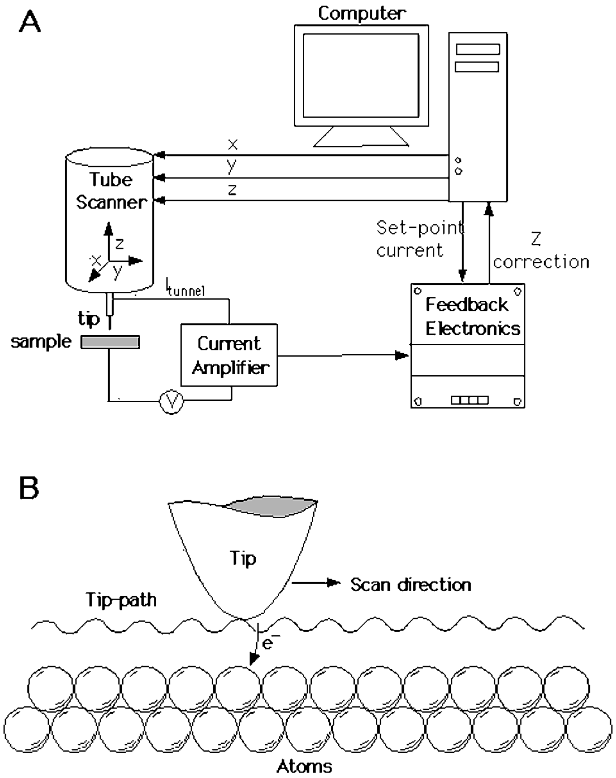


FIGURE 1 (A) Schematic illustration of a scanning tunneling microscope. (B) Tip trajectory as it tracks surface atomic structure in constant-current mode.

measured current is higher than the setpoint current, the feedback electronics applies voltages to the Z axis of the tube scanner to make the tip retract until the current equals the setpoint. Conversely, if the measured current is too small, voltages are applied to the tube scanner to extend the tip closer to the substrate until the current increases to the setpoint value. Essentially, the goal of the feedback electronics is to zero out the "error" signal, or the difference between the measured current and the setpoint. This electronic feedback process is very quick, requiring only fractions of a second, so that the tip may track the surface topography while scanning at speeds from 100 to several thousand nanometers per second.

The tube scanner is calibrated and therefore it is known how many angstroms the tip moves in the X, Y, and Z directions for given applied voltages. Topographic height images are produced by recording the Z-axis position of the tip as a function of its X and Y coordinates (pixel position). It is a remarkable fact that the constant-tunneling-current methodology allows the tip to track the topographic contours of individual atoms, as shown schematically in Fig. 1B. The resolution of STM height images is typically better than 1 Å in the X and Y directions and 0.1 Å in the Z direction, sufficient to visualize individual atoms. This superb resolution is achieved because the tunneling current is emitted essentially from the last atom on the end of the tip. In reality, tips are not smooth as indicated by Fig. 1B, but have atomic corrogation and many different asperities or "minitips" that can emit tunneling current. However, because the tunneling current (in vacuum) decreases by about a factor of 10 for every 1 Å of separation, and because generally there is one minitip that is a few angstroms closer to the sample than the others, it is usually the case that the vast majority of the tunneling current flows through only one protrusion. It is really the exponential distance dependence of tunneling current that gives STM its tremendous resolution.

It is important to note that the magnitude of the tunneling current depends not only on the tip–sample separation, but also on the electron density of states (number of states per volume per unit energy) in the sample and the tip bias. If the tip is biased negatively, electrons tunnel from the tip into empty electronic states in the sample surface. If the tip is biased positively, electrons tunnel from filled states in the sample to the tip. Therefore, the contrast or corrogation seen in STM images is really a map of the spatial distribution of electron states, either empty or filled, depending on the tip bias. In many cases, the electron density directly reflects atomic positions on the sample. As we shall see in Section II, however, this is not always the case. Additionally, there can be a nonlinear dependence of the tunneling current on the tip voltage, reflecting the spectrum of the electronic energy states at the sample surface. Examining the voltage dependence of the tunneling current is the basis of local tunneling spectroscopy, which is not covered in this article, but is discussed extensively by Wiesendanger (1994).

STM imaging can be accomplished in air, vacuum, or under fluids, though the best quality images are usually taken under vacuum since contamination of the tip or sample by adsorbates (e.g., water) is minimized. The time required to produce an atomic-resolution STM image is tens of seconds to minutes depending on the scan size and scan speed. Samples for STM analysis must be conductive.

B. Atomic Force Microscopy (AFM)

Atomic force microscopy uses the repulsive or attractive forces between the probe and sample rather than tunneling current to track sample topography. Figure 2A shows an

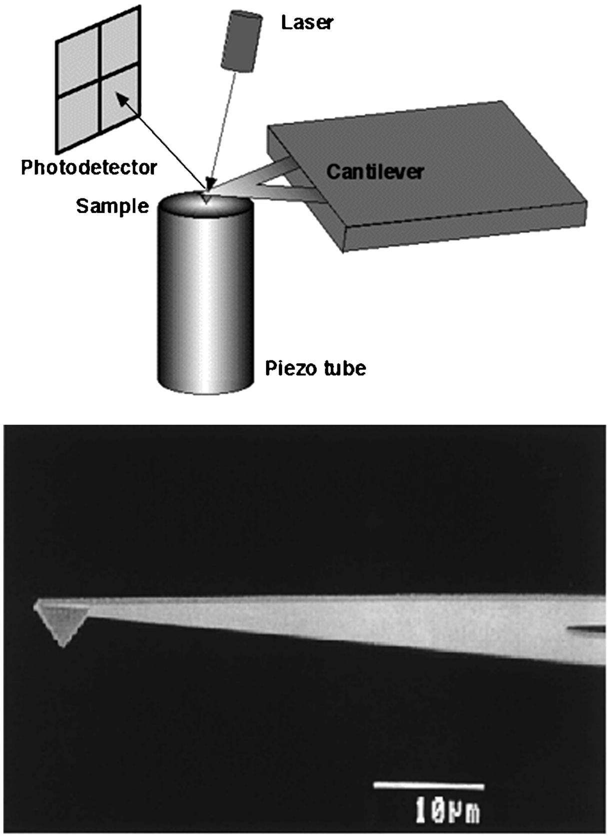


FIGURE 2 (A) Schematic illustration of atomic force microscopy. Cantilever deflection recorded at the photodetector is used to generate a Z-position feedback signal, analogous to the way tunneling current is used to control Z position in STM. (B) Scanning electron micrograph of a commercial silicon nitride AFM tip and cantilever. [Reprinted with permission from Takano, H., Kenseth, J. R., Wong, S.-S., O'Brien, J. C., and Porter, M. D. (1999). *Chem. Rev.* **99,** 2845–2890. Copyright 1999 American Chemical Society.]

illustration of a typical AFM apparatus. The probe, shown in Fig. 2B, consists of a sharp tip on the end of a small cantilever a few hundred micrometers in length. Forces exerted on the tip cause the cantilever to bend. The force is directly proportional to the cantilever deflection and is given by

$$F = k\Delta z,$$

where k is the cantilever force constant (in nN/nm, analogous to a spring constant) and Δz is the vertical deflection. Cantilevers can be fabricated with force constants as low as 0.01 nN/nm, so that forces as small as 0.1 nN give nanometer deflections that can be detected easily. The sensitivity of AFM to forces smaller than the bonding forces between individual atoms was the motivation for the name “atomic force” microscopy. Some investigators instead use the name scanning force microscopy (SFM) to emphasize the relationship to other scanning probe techniques.

In the simplest version of AFM, called contact mode or normal force mode, the tip is brought into hard contact with the sample surface such that the cantilever bends due to the tip–sample repulsion (Fig. 3A). The cantilever deflection can be measured in several ways, but the most common is the optical beam method. In this approach, a laser beam from a semiconductor laser is focused on the back side of the cantilever, as shown in Fig. 2A. The reflected light hits a position-sensitive detector that produces a voltage indicating how much the cantilever has deflected up or down. In some cases, the detector also produces a signal indicating how much the cantilever has deflected sideways. The optical beam method allows deflections as small as an angstrom to be detected.

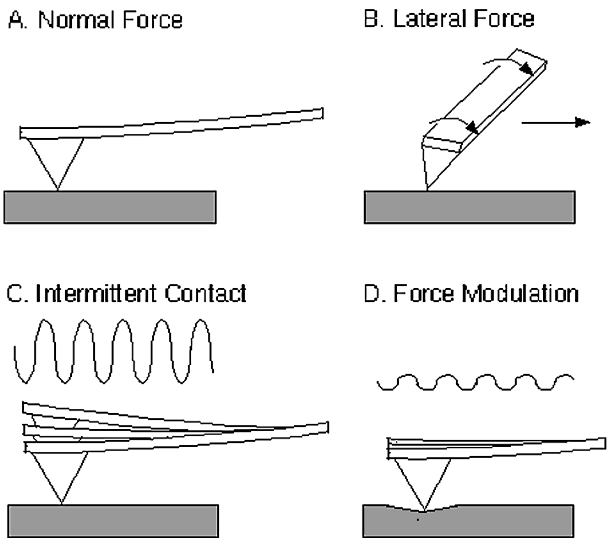


FIGURE 3 Different AFM imaging modes. (A) Contact or normal force mode. (B) Lateral or friction force mode. (C) Intermittent contact mode, also known as noncontact or tapping mode. (D) Force modulation mode.

AFM instruments are designed to scan either the sample or the cantilever-tip assembly. Figure 2A shows a sample scanning configuration. In contact mode operation, the sample is scanned underneath the cantilever tip, which remains stationary. The user defines a setpoint cantilever deflection (actually, a voltage proportional to the deflection) that is to be maintained by the feedback electronics during scanning. If the cantilever deflection increases, the feedback electronics applies a voltage to the Z axis of the piezo to cause the sample to retract from the lever, thus lowering the deflection back to the setpoint value. Conversely, if the cantilever deflection decreases, the feedback electronics applies a Z-axis voltage to cause the sample to move toward the lever, increasing the deflection back to the setpoint value. As in STM, the feedback is done very quickly, and recording the Z position as a function of X–Y coordinates produces a topographic map or height image.

Resolution in AFM is governed principally by the sharpness of the probe tip, although the sample characteristics and specific operating mode are also important. Commercial cantilever-tip assemblies are made from silicon or silicon nitride using conventional microfabrication methods common to the semiconductor industry. Figure 2B shows an electron micrograph of a microfabricated silicon nitride probe. Commercial probes can be made with tip radii as small as 10 nm, but 20–40 nm is more common. With a 20-nm-radius tip it is possible to obtain images of surfaces that show topographic corrogation due to atomic or molecular packing.

As AFM has matured, a number of more sophisticated imaging modes have been developed. In lateral force mode (Fig. 3B) the sample is scanned sideways relative to the long axis of the cantilever. Torsional forces exerted on the tip cause the cantilever to twist and consequently to deflect the optical beam horizontally on the photodetector. Recording the lateral deflection or twisting of the cantilever as a function of X–Y position gives a lateral force map. Lateral force images are particularly sensitive to friction force between the tip and sample, and therefore these images are also called friction force maps.

Two other popular scanning modes involve oscillating the cantilever-tip assembly while the sample is scanned (Figs. 3C and 3D). In tapping mode (Fig. 3C) the tip is oscillated vertically with an amplitude of 50 to several hundred nanometers and makes contact with the sample at the bottom of every downstroke. The optical beam and photodetector assembly is used to detect the oscillation amplitude, which is dependent on how close the tip is to the sample. The feedback control electronics maintains the oscillation amplitude at a user-defined setpoint

value and a topographic image is generated in the same fashion as in contact mode, i.e., by recording the Z position of the sample as a function of X and Y coordinates. The advantage of this intermittent contact method is that shear forces that are always exerted on the sample in contact mode are eliminated. For imaging soft samples, such as biological specimens, this has proved to be an advantage. If the tip is allowed to oscillate near the surface but not to touch it (called noncontact mode), it is possible to obtain true atomic resolution under favorable circumstances.

Another operation mode is known as the force modulation method. This is a contact mode technique, but instead of maintaining fixed cantilever deflection, the tip is periodically pushed into the sample, as shown in Fig. 3D. This periodic force modulation is accomplished by oscillating the Z position of the sample at a known amplitude (typically 2–5 nm). If the region under the tip is stiff, the probe will not compress the sample and the cantilever will deflect by the oscillation amplitude (e.g., 2 nm). If the sample is soft, the probe will push into the sample and the cantilever will deflect somewhat less than the full oscillation amplitude. Thus, recording the modulation in tip deflection as a function of X and Y position produces a stiffness map of the sample.

Most AFM techniques (e.g., contact, lateral force, and tapping modes) can be performed in vacuum, air, or under liquids. AFM can be applied to insulators as well as conductors, which means the technique is applicable to a wider range of materials than STM.

II. APPLICATIONS OF SCANNING TUNNELING MICROSCOPY (STM)

Since its invention in 1981, the STM has been valued for the remarkable atomic-resolution images it produces. An excellent example is shown in Fig. 4, which shows a 10 nm × 10 nm topographic image of the surface of a silicon crystal. The bright spots in the image correspond to individual silicon atoms that are 2.4 Å in diameter, highlighting the superb resolution of the microscope. Indeed, it is an attractive feature of STM images that atomic-level defects can be resolved; there are clearly atoms missing (vacancies) in some portions of the image. Historically, the STM observation of the Si "7 × 7 reconstruction" was extremely important since the atomic arrangement of this surface was in dispute until the atomically resolved STM images became available.

The exquisite imaging capabilities of STM have been applied to a wide range of surface chemistry and physics problems. Figure 5 shows a 40 nm × 40 nm image of a silicon surface that was exposed to bromine vapor (Br_2) at 800 K. At this high temperature, bromine reacts with

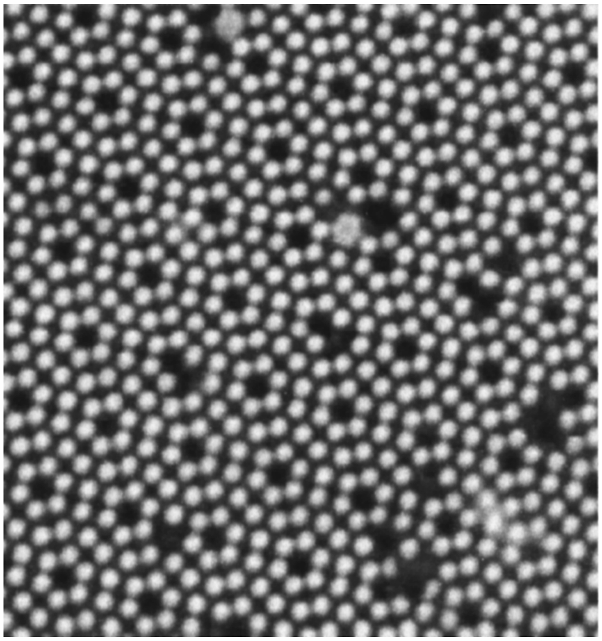


FIGURE 4 A 10 nm × 10 nm atomic-resolution STM image of the surface of a silicon crystal. This image was taken on the (111) face of the crystal, which exhibits an unusual atomic reconstruction (the "7 × 7" reconstruction). Individual atoms and defects are observable. The image was acquired in constant-current mode with a current of 2 nA and a tip voltage of +1.99 V. [Courtesy of Omicron Vakuumphysik GmbH.]

silicon to form silicon bromide ($SiBr_4$), which is also a vapor. Thus, this reaction results in etching of the silicon surface. The atomic-resolution STM image shows how this etching occurs. Two planes (or terraces) of silicon atoms are visible in which pits have formed. These pits are elongated along rows of silicon atoms for reasons related to the energetics of etching. In addition, pits on the top plane (upper left) are oriented at 90° with respect to the bottom plane (lower right) because of details of the silicon crystal structure. Also evident are bright lines on top of the terraces that run perpendicular to the pits. These lines represent regrowth islands, produced from silicon atoms that are released onto the terrace during an etching event. STM is a useful tool in understanding such etching processes with atomic detail.

STM is also used to study molecular adsorption to surfaces, a process that is important in the growth of thin films and the surface catalysis of chemical reactions. Figure 6 shows STM images of two different phthalocyanine molecules adsorbed to a gold crystal. As the figure shows, phthalocyanine molecules are flat, cloverleaf structures about 1 nm across and contain a complexed metal ion in the center. The images clearly show the shape of the individual phthalocyanine molecules and also reveal that they form an ordered overlayer on the gold crystal

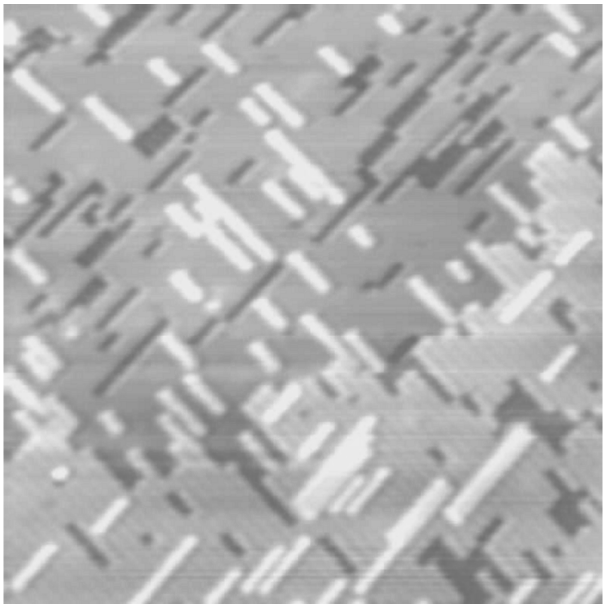


FIGURE 5 A 40 nm × 40 nm STM image of a silicon surface that has been exposed to a halogen etchant (bromine vapor, Br_2). Terraces of silicon atoms are visible containing elongated etch pits (dark stripes) and regrowth islands (bright stripes). Pits on the upper and lower terraces are oriented perpendicular to each other because of the silicon crystal structure. [Courtesy of John H. Weaver, University of Illinois.]

substrate. A particularly interesting observation is that the nickel ions at the center of the molecules on the left show up as holes (dark spots), whereas the iron atoms in the molecules shown on the right appear as peaks (bright spots). This is not because there are actual height differences in the two images, because both molecules are planar. Rather, the interpretation rests on the different electronic energy states in the iron- and nickel-containing molecules. In these experiments, the tip was biased negatively with respect to the samples; consequently, the images reflect the density of empty states in the molecules. Iron(II) ion has fewer electrons than Ni(II) ion and consequently there are more empty states (*d* orbitals) to tunnel through, giving iron brighter contrast. This is an excellent example of how, through its sensitivity to electronic structure, STM detects chemical differences in adsorbed molecules as well as topographic differences. In a similar fashion, a number of workers have shown that STM images can distinguish double bonds from single bonds in adsorbed molecules.

In addition to its use in studying surface structure, surface reactivity, and molecular adsorption, STM is also finding use in nanofabrication, or the making of extremely small objects. One outstanding example of the use of STM as a fabrication tool is shown in the series of images in Figure 7, which depict the assembly of a ring of 48 iron atoms on a copper single-crystal surface. This process was carried out in an STM enclosed in ultrahigh vacuum with the copper surface held at 5 K. A submonolayer quantity of iron atoms was deposited on the copper. The STM tip was used to visualize these adsorbed atoms and to drag them across the surface to form the ring structure. The key to this process is the ability to grab an atom with the STM tip. This can be done by applying a voltage pulse to the tip such that the iron atom hops off the surface and onto the tip. The iron atom can then be redeposited somewhere else by applying another voltage pulse.

Perhaps the most remarkable aspect of the images in Fig. 7 is the observation of electron standing waves inside the completed ring. Electrons are quantum mechanical in nature, meaning they exhibit wavelike properties. Confining an electron within a two-dimensional "corral" is predicted to yield a circular electron density wave. Figure 7 shows this wave is visualized directly by STM of the completed corral. This is a particularly beautiful confirmation of the wavelike properties of electrons. Like the images in Fig. 6, the visualization of this electron wave is possible because STM is sensitive to electron density. Note that at the bias voltage chosen, the copper atoms underneath the ring are not readily apparent.

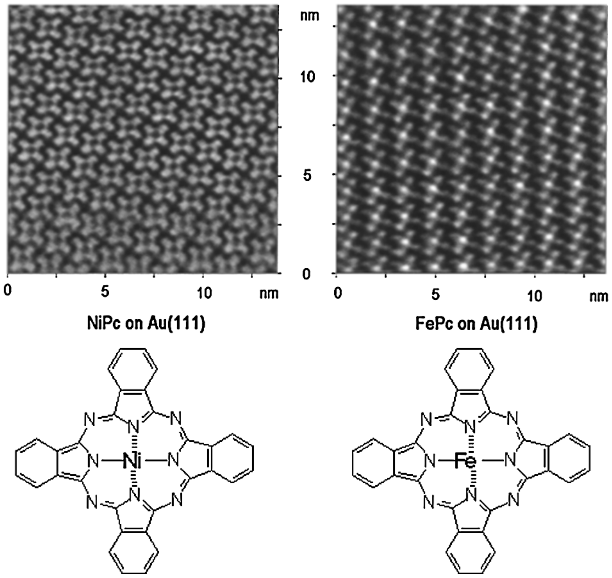


FIGURE 6 STM images of disc-shaped phthalocyanine molecules on single-crystal gold. Molecular structures are shown below the images. The image on the left shows nickel-containing phthalocyanines ordered on the gold surface. The image on the right shows iron-containing phthalocyanines. Note that the nickel ions (Ni^{2+}) on the left appear as holes (dark spots) and the iron ions (Fe^{2+}) on the right appear as peaks (bright spots) due to differences in their electronic structure. [Reprinted with permission from Lu, X., and Hipps, K. W. (1997). *J. Phys. Chem. B* **101,** 5391–5396. Copyright 1997 American Chemical Society.]

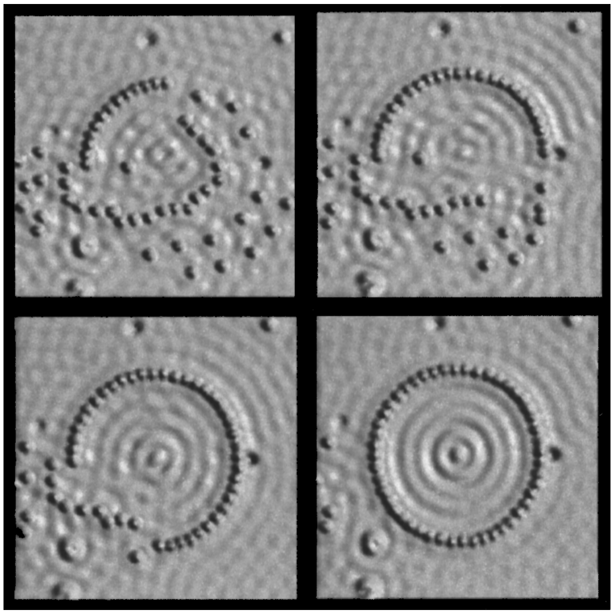


FIGURE 7 A series of STM images showing the assembly of a ring of 48 iron atoms on a copper surface. The ring diameter is 14.3 nm. The assembly and imaging was accomplished in ultrahigh vacuum (10^{-10} torr) using a tungsten tip with the sample held at 5 K. The image in the lower right shows a circular standing electron density wave confined within the completed ring. Tip current 1 nA; tip voltage 0.01 V. [Courtesy of Don Eigler, IBM Almaden Research Center.]

III. APPLICATIONS OF ATOMIC FORCE MICROSCOPY (AFM)

Like STM, AFM can be used to view the geometrical arrangement of atoms or molecules on surfaces, and the technique can be applied to electrical insulators as well as conductors. One important application of AFM is the visualization of crystallization processes in solution. Figures 8A and 8B show contact mode topographic images of the surface of a growing insulin crystal. These images were acquired under an aqueous solution containing dissolved insulin. Terraces are visible that correspond to single layers of insulin hexamers. Figure 8C shows a molecular-resolution image taken in contact mode on one of the terraces. The molecular packing symmetry observed can be used to determine which crystal face is viewed; Fig. 8D shows a computer representation of the packing of insulin hexamers on the (001) crystal face.

Increasingly, AFM is being applied to biological imaging and visualization of individual macromolecules such as DNA. Figure 9A shows tapping mode images taken under solution of human chromosomes. Figure 9B shows a 500 nm × 500 nm tapping mode image of closed-loop (bacterial plasmid) DNA molecules adsorbed to a mica surface. It should be appreciated that the DNA molecules appear much wider than they actually are because the tip is not infinitely sharp but has a finite radius, typically 20 nm. Consequently, the observed width represents a convolution of the actual dimensions of the DNA with the tip shape. However, the thickness of the DNA chain can be accurately estimated by measuring the height of the DNA molecule because the height is not affected by the tip shape. By this method, the thickness of the DNA molecules is estimated to be about 1 nm (10 Å).

AFM is also used as a tool to investigate frictional and mechanical properties of thin films on the nanoscale. Figure 10 shows topography, friction, and elasticity (or stiffness) maps of an organic film composed of hydrocarbon and fluorocarbon molecules. These three images were acquired simultaneously in air by recording the average Z position, average lateral cantilever deflection, and the modulated vertical deflection as the tip was scanned over the surface in force modulation mode. The scan direction

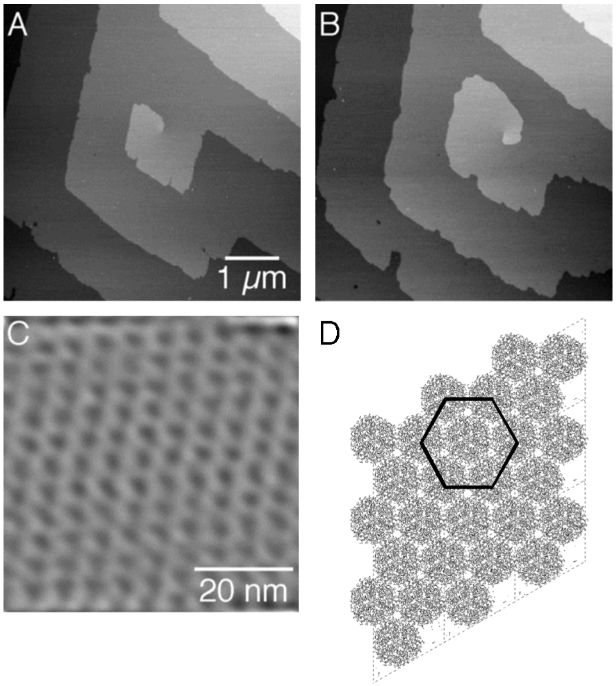


FIGURE 8 (A, B) Contact mode AFM images of the crystal growth of genetically engineered insulin recorded approximately 3 hr apart under a phosphate-buffered saline solution containing the insulin. The step heights of the terraces correspond to a single layer of insulin hexamers. (C) A contact mode lattice image acquired on one of the terraces during growth, revealing the hexagonal packing of the insulin hexamers. (D) Illustration of the packing of the insulin hexamers, viewed normal to the (001) plane of the rhombohedral crystalline form. [Courtesy of Michael D. Ward, University of Minnesota.]

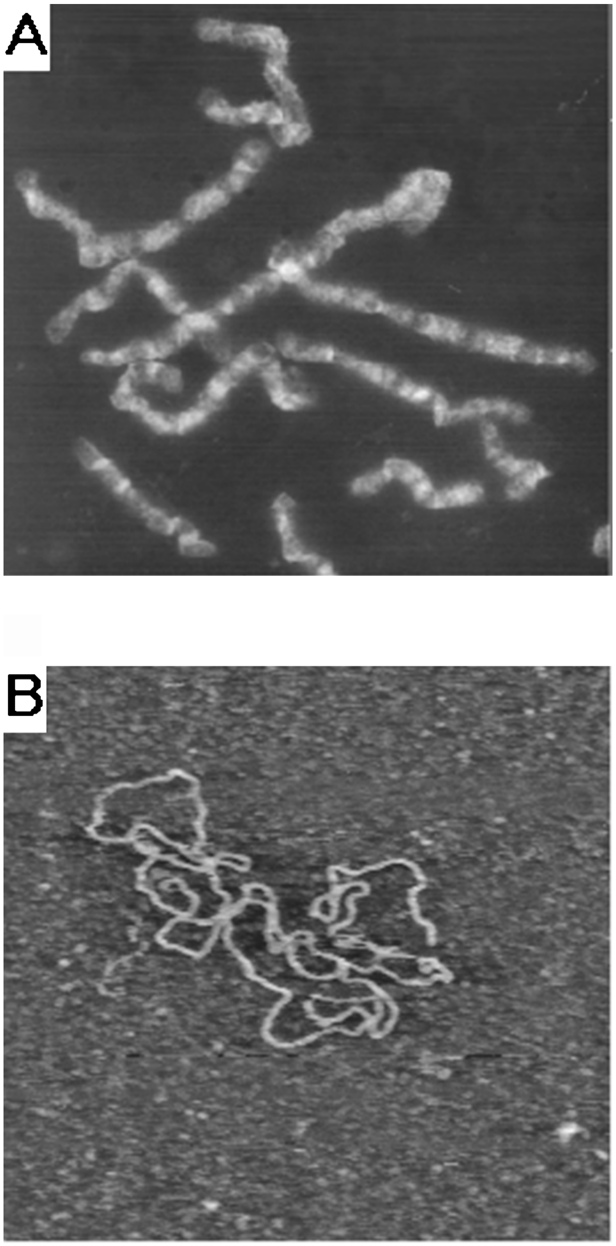


FIGURE 9 Tapping mode AFM images of human chromosomes (A) and plasmid DNA (B), taken under aqueous buffer solution. The image in (A) is 25 μm $\times$ 25 μm. The image in (B) is 500 nm $\times$ 500 nm. [Courtesy of Digital Instruments, Inc.]

was perpendicular to the long axis of the cantilever, enhancing sensitivity of the instrument to frictional forces. In these mixed-molecule films, the hydrocarbon molecules assemble into 50- to 100-nm-diameter islands on top of a sea of fluorocarbon. The height image on the left shows that the islands protrude from the surface. Importantly, comparison of the lateral force image in the center with the elasticity image on the right shows a connection between friction and sample stiffness or compressibility. The friction image shows that higher friction is observed on the fluorocarbon areas (which are bright) than on the hydrocarbon islands (which are dark). The elasticity map on the right has the opposite contrast and shows that the hydrocarbon is less compliant than the fluorocarbon. It can be concluded, therefore, that dissipative frictional forces correlate with the ease with which the film can be compressed. Such fundamental information is useful to investigators interested in the molecular mechanisms of friction and lubrication.

Nanometer-scale mechanical analysis of surfaces, films, and even individual molecules can also be accomplished by recording cantilever deflection as a function of Z position. Because the lever deflection is easily converted to force, these measurements are called "force curves" or F–Z measurements, and are shown in Fig. 11. The tip is held at a fixed X–Y coordinate and the Z position is varied. As the sample gets very near the tip, attractive van der Waals forces cause the tip to jump to hard contact with the surface (position 2). Further extension of the sample toward the cantilever results in deflection of the lever upward (position 3) and an increasing compressive force exerted on the tip–sample microcontact due to the lever deflection. Reversing the sample direction allows examination of the adhesive interaction between the tip and sample. When the sample retracts to position 4, the cantilever deflection is zero, but the tip and sample do not separate because of adhesion. Continued retraction of the sample produces downward deflection of the lever and a tensile force on the tip–sample microcontact. Only when this tensile force just exceeds the adhesion force will the tip pop off the sample (position 5). Noting the cantilever deflection at this rupture point allows calculation, via the lever spring constant, of the adhesion or pull-off force. Such pull-off measurements are used to examine adhesion in local regions on surfaces. These measurements also are used by researchers to measure nanomechanical behavior and forces associated with stretching *individual* long-chain molecules, like proteins and DNA.

IV. OTHER SPM TECHNIQUES AND APPLICATIONS

A. Near-Field Scanning Optical Microscopy (NSOM)

In NSOM, a light-emitting tip is scanned over a sample surface to produce a high-resolution topographic and optical image simultaneously. The topographic image is generally obtained using some type of force feedback

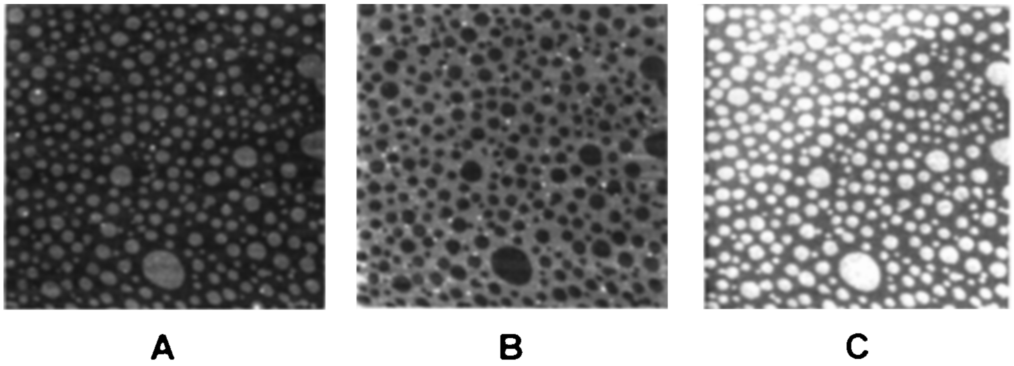


FIGURE 10 Simultaneously acquired topography (A), lateral force (B), and elasticity (C) maps of a phase-separated hydrocarbon and fluorocarbon film taken by atomic force microscopy. Each image is 3 μm × 3 μm. The raised (bright) islands in the topographic image contain the hydrocarbon molecules that lie on top of a sea of fluorocarbon. The lateral force map shows that the hydrocarbon islands, which appear dark, exhibit lower friction. The elasticity map has the opposite contrast and shows that the islands are less compliant (or stiffer) than the fluorocarbon regions. These images allow correlation of friction with mechanical compliancy. [Reprinted with permission from Overney, R. M., Meyer, E., Frommer, J., Guntherodt, H.-J., Fujihira, M., Takano, H., and Gotoh, Y. (1994). *Langmuir* **10,** 1281–1286. Copyright 1994 American Chemical Society.]

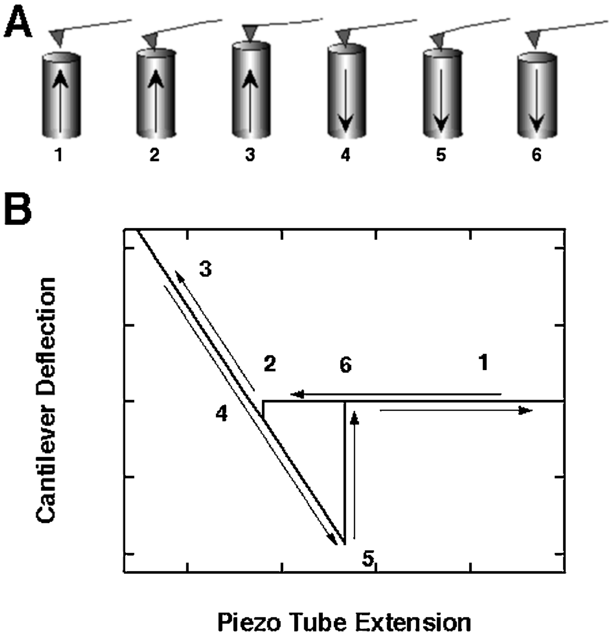


FIGURE 11 (A) Scheme of force versus Z position (F–Z) measurements. (B) Cantilever deflection (which can be converted to force) is plotted as a function of Z position (or piezotube extension). After the tip has made contact with the sample (just after position 2), continued extension of the piezo results in increasing deflection of the lever. Reversing the piezo direction allows measurement of the tip–sample adhesion force. When the piezo has retracted to position 5, the cantilever is now bent downward, exerting a tensile force on the tip–sample microcontact. Further retraction results in tip pull-off (5 to 6). The cantilever deflection at position 5 allows estimation of the adhesion force using the formula $F = k(\Delta z)$, where Δz is the downward deflection of the cantilever and k is the cantilever spring constant.

discussed below. In illumination mode, the optical image is acquired by feeding light into the probe and collecting the transmitted, reflected, or fluorescent emission from the sample with an objective lens. Light intensity at each pixel point (X–Y coordinate) is recorded by a photodetector. In the relatively less common collection mode, illumination is provided by an external source and the *probe* is used to collect transmitted, reflected, or fluorescence emission from the sample at each pixel. Figure 12 shows an illumination mode setup with the collection objective and tip on opposite sides of the sample; other designs have the collection optics and probe on the same side. Spatial resolution of about 50 nm is achievable routinely in NSOM images, which is about 10–20 times better than the diffraction-limited resolution in conventional light microscopes.

NSOM probes are typically made from optical fiber that has been heated and tapered in a micropipet puller. An aluminum coating is evaporated onto the sides of the probe leaving an open aperture at the tip 10–100 nm in diameter, as illustrated in Fig. 13A. Laser light coupled into the fiber propagates to the tip, where it emerges from the aperture. The aluminum cladding confines the light in the X and Y dimensions to the diameter of the aperture, and thus spatial resolution is determined by the size of this opening. Unfortunately, the optical power of the probe tip also depends very sensitively on the aperture size. The throughput of the probe drops dramatically as the aperture size is decreased, and thus there is a tradeoff between resolution and sensitivity. For a good probe with a 50-nm aperture, 0.5 mW coupled into the fiber results in greater than 1 nW of radiation from the tip.

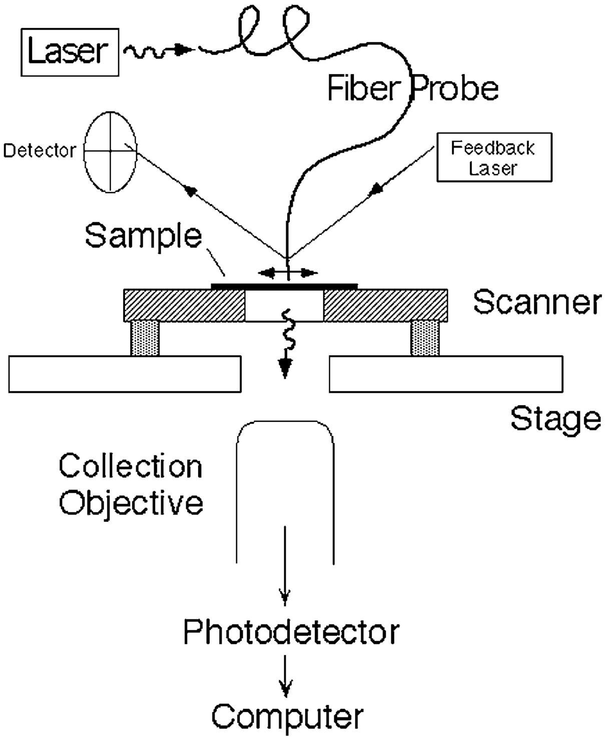


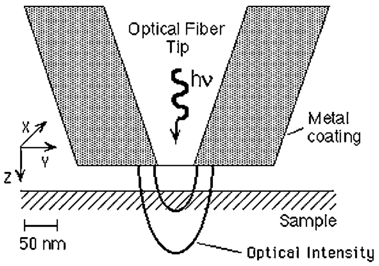
FIGURE 12 Schematic illustration of an NSOM setup in illumination mode. Laser light coupled into the optical fiber propagates to the probe tip, where it interacts with the sample. Transmitted or fluorescence emission is collected below the sample stage by an objective lens, which focuses the light on a photodetector. A computer records the light intensity as a function of *X*–*Y* coordinates. Shear force *Z*-position feedback, as described in the text, is used to keep the probe close to the sample and to track the sample topography. Often NSOM instruments are mounted on the stage of a conventional light microscope.

NSOM is a so-called near-field technique because the optical radiation near the tip (the "near-field") is orders of magnitude greater than the radiation just 10 nm away from the aperture. Models predict that the radiation decays almost exponentially with increasing distance from the probe. Consequently, it is important that the probe be scanned with the aperture as close as possible to the sample surface to maximize contrast and resolution. Flat, smooth tip ends have been shown to improve image quality vastly by allowing for closer proximity of the aperture to the sample. Tip ends can be flattened and polished by focused ion beam (FIB) milling. Figure 13B shows examples of two ideal NSOM tips after FIB milling.

To keep the aperture near the sample and to track the sample topography, shear force feedback is commonly employed. In this method, a straight fiber probe with its long axis perpendicular to the substrate is oscillated (or "dithered") by a piezoelectric device such that the aperture moves side to side. This is in contrast to the tip vibration modes discussed in Section I, where the tip oscillates up and down. As the vibrating fiber probe approaches the sample, shear forces between the probe and tip dampen the oscillation amplitude. The shift in amplitude or frequency is detected and used to generate a feedback signal to the *Z*-position piezo. The probe vibration can be detected optically or piezoelectrically using a tuning fork in which the fiber probe is attached to one arm. Whatever the method of detection, the upshot is that the probe vibration is kept constant by the feedback electronics. Using this shear force method, topographic images can be generated simultaneously with the near-field image.

NSOM can produce high-resolution transmission and fluorescence images of a variety of samples including polymer films, proteins, chromosomes, molecular crystals, and even isolated molecules. Figure 14a shows a fluorescence NSOM image of individual dye molecules in a polymer matrix. The molecules are much smaller than the tip aperture diameter, so that the fluorescent spots observable in the image are essentially maps of the electric field distribution near the tip. The diameter of these spots is an estimation of the probe resolution ($\sim$120 nm in this case).

A



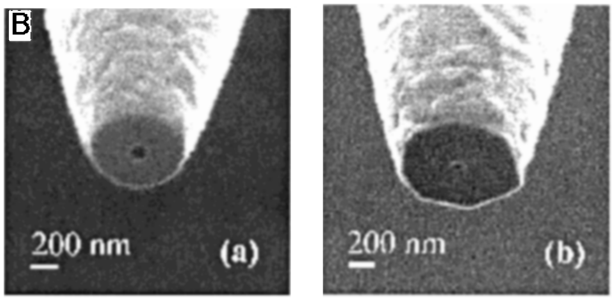


FIGURE 13 (A) Schematic illustration of an optical fiber tip used for NSOM. The fiber is drawn to a sharp point and coated with a metal, such as aluminum, to keep the light confined to the fiber core. Radiation emerges from the small aperture at the end of the tip and interacts with the sample. The radiation intensity decreases rapidly with distance in the *Z* direction. (B) Electron micrographs of aluminum-coated NSOM tips that have been flattened and polished by focused ion beam milling. [Reprinted with permission from Veerman, J. A., Otter, A. M., and van Hulst, N. F. (1998). *Appl. Phys. Lett.* **72,** 3115–3117.]

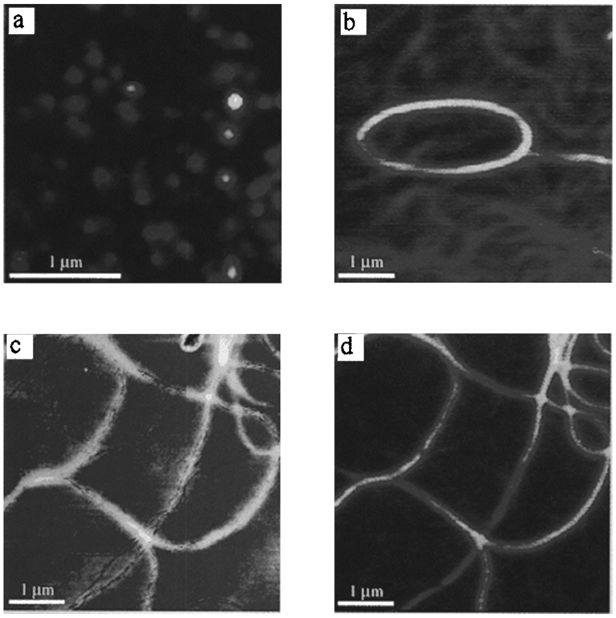


FIGURE 14 (a) Fluorescence NSOM image of individual dye molecules embedded in a thin film of poly(methyl methacrylate) supported on a quartz coverslip. The fluorescent regions of the film result from excitation and detection of single, isolated dye molecules. The molecules are much smaller than the tip aperture diameter, so that the fluorescent spots are essentially images of the electric field distribution near the tip. The diameter of these spots is an estimation of the probe resolution (~120 nm in this case). (b) Fluorescence NSOM image of a thin polymer film containing imbedded fluorescent fibers. (c, d) Topography and fluorescence NSOM images of fibers in a similar sample as in (b). [Reprinted with permission from Higgins, D. A., Kerimo, J., Vanden Bout, D. A., and Barbara, P. F. (1996). *J. Am. Chem. Soc.* **118,** 4049–4058. Copyright 1996 American Chemical Society.]

Figures 14b–14d show topography and NSOM images of flexible fluorescent fibers imbedded in a thin polymer film. It is also possible to map birefringence and dichroism in samples using polarized-light NSOM. More recently, there has been increasing interest in imaging opaque samples by collecting reflected light. An interesting example of this is the imaging of magnetic films using polarized radiation. Due to the magnetooptical Kerr effect, polarized-light NSOM images reveal magnetic domains.

An additional attractive feature of NSOM is that it can be used to perform localized spectroscopy. For example, the NSOM probe can excite the dye aggregates in Fig. 14 at a particular location and the spectrum of the fluorescence emission can be recorded by sending the collected light to a spectrophotometer. The possibility of localized spectral analysis of sample emission dramatically increases the amount of spatially resolved, detailed structural and chemical information that can be determined by NSOM. Time-resolved NSOM spectral studies have also recently been used to investigate dynamic processes (e.g., charge transfer reactions) on submicrometer scales.

B. Magnetic Force Microscopy (MFM)

Magnetic Force Microscopy (MFM) is a noncontact scanning force technique that uses magnetostatic interactions between tip and sample to image magnetic domains. MFM is a derivative of atomic force microscopy and can be performed with many commercial AFM instruments. The magnetic structure of a wide variety of samples can be investigated with MFM with a resolution of up to 10 nm, depending on the sharpness of the probe.

MFM probes are often prepared by sputtering a ferromagnetic film such as cobalt–nickel or iron onto a conventional silicon microfabricated cantilever/tip assembly. Alternatively, a magnetic probe can be fashioned out of an etched, bent magnetic wire. To produce an image, the magnetic tip is oscillated at the cantilever resonance frequency and scanned over the sample at a distance of at least a few nanometers *above* the surface. During the scan, the tip does not touch the sample, but magnetic fields at the surface interact with the magnetic tip. Gradients in the magnetic field near the sample surface modify the effective spring constant of the cantilever, thereby shifting its resonant frequency. The shift in frequency and the accompanying change in phase and amplitude of the cantilever oscillation are detected using an optical sensor as described above for AFM. Plotting the frequency shift, amplitude, or phase of the cantilever as a function of sample position creates an MFM image.

Because MFM detects magnetic field *gradients*, MFM images highlight boundaries between magnetic domains where the north–south dipoles are pointing in different directions. The sensitivity to magnetic field gradients also means MFM images are strongly influenced by sample topography. One approach to removing or mitigating "topographic contrast" in magnetic force images is to perform a double scan. First, a topographic scan line is recorded (e.g., in tapping mode) and stored to memory. Then the probe is lifted off the surface by a few nanometers and the scan line is repeated, except this time the magnetic force signal is recorded while the probe tracks the "memorized" sample topography. The double-scan mode allows acquisition of both topographic and magnetic images and is an effective way to deconvolute surface topography from magnetic information.

MFM is useful for imaging all kinds of magnetic materials, including technologically important data storage materials such as magnetic and magnetooptic recording media. For this application, an attractive feature of MFM is that the domains can be imaged through nonmagnetic and opaque overcoats that are often applied to information

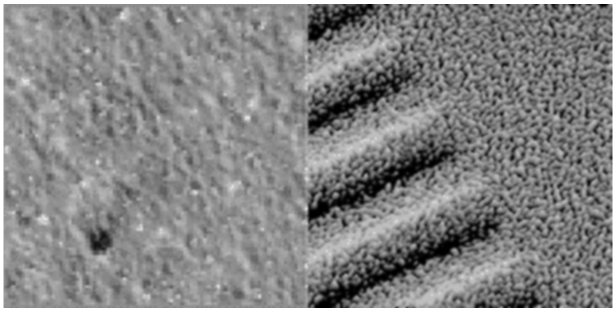


FIGURE 15 Two 12 μm × 12 μm topographic (left) and magnetic force (right) images of a track edge on magnetic recording tape. The magnetic force image was taken at a lift height of 30 nm. The bright and dark regions in the magnetic force image correspond to boundaries between the magnetic bits. [Courtesy of Digital Instruments.]

storage materials. Figure 15 shows topographic and magnetic force images of a track in magnetic recording tape. The bright regions of the magnetic force image correspond to the boundaries between magnetic bits in which the north–south dipoles are oriented in opposite directions. Note that the simultaneously acquired topographic image shows very little structure, i.e., the tape is topographically flat. Images of the magnetic track give information about the bit shape, position, and edge roughness. MFM is also used extensively in magnetic materials research, for example, to study the pinning of magnetic domain walls by structural features such as grain boundaries.

C. Electric Force Microscopy (EFM) and Scanning Surface Potential Microscopy (SSPM)

Electric force microscopy (EFM) is exactly analogous to MFM with the exception that the tip is used to sense electric field gradients instead of magnetic field gradients. EFM requires conducting AFM tips that interact with the sample through long-range Coulomb forces. The probes can be microfabricated from doped silicon or they can be made by metal coating silicon or silicon nitride cantilever/tip assemblies. As in MFM, the tip is oscillated at the cantilever resonance frequency and scanned over the sample at a distance of a few nanometers above the surface, such that the probe does not make physical contact with the sample. Electric field gradients associated with charges or electric dipoles present on (or slightly beneath) the surface interact with the tip and cause shifts in the cantilever resonance frequency. EFM images are created by recording the cantilever amplitude, phase, or resonance frequency shift as a function of sample position. Contrast in EFM images can be manipulated by applying dc voltages to the conducting tip to enhance attractive or repulsive electric forces. As in MFM, topographic effects may be removed by first scanning the sample topography, then lifting the tip and rescanning at a constant height (constant tip–sample separation) while acquiring the EFM signal. In principle, EFM images can be quantitatively related to charge distributions on a sample surface through mathematical modeling of the tip–sample interaction.

Figure 16 shows 80 μm × 80 μm topographic and EFM images of an active integrated circuit. The light areas of the topographic image highlight the location of aluminum wires. The EFM image on the right shows which of the wires have an applied voltage.

Scanning surface potential microscopy (SSPM) is a related electrical characterization method that produces maps of electrostatic potential across a sample surface. SSPM uses the principles of the Kelvin probe method, which is a nulling technique, and is therefore also known as Kelvin-probe force microscopy (KFM). In contrast to EFM, during the lift scan the tip is not vibrated mechanically. Instead, an oscillating voltage, with a frequency equal to the resonant frequency ω of the cantilever, and a variable dc offset voltage are applied to the cantilever and tip. As the tip travels above the sample surface in lift mode, the tip and the cantilever experience an oscillating force proportional to the square of the potential difference between the tip and the sample ($\Delta V = V_{\text{tip}} - V_{\text{sample}}$). The electronics detects the lever oscillation at frequency ω and adjusts the dc bias on the tip such that the oscillation amplitude goes to zero. When the adjustable dc tip bias V_{tip} equals the surface potential (i.e., $V_{\text{tip}} = V_{\text{sample}}$), the cantilever experiences no oscillating force at frequency ω, and V_{tip} is identical to the surface potential at that position on the sample. SSPM images consist of maps of V_{tip} (i.e., surface potential) as a function of X–Y coordinates. The approach has enhanced signal-to-noise ratio common to frequency lock-in and signal nulling techniques and

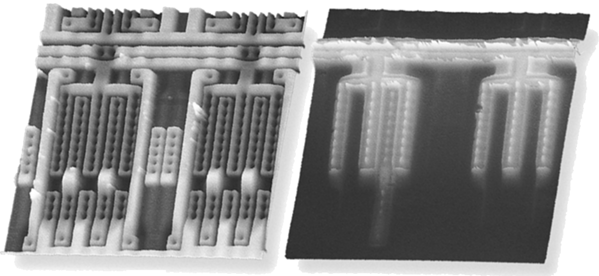


FIGURE 16 Topography (A) and EFM image (B) of a live packaged integrated circuit (IC) with a passivation overlayer. Displayed area is 80 μm × 80 μm. The EFM image shows which metal wires have voltages applied to them. The bright region between two of the wires in the three-pronged fork on the left indicates a transistor in saturation (fully on). [Courtesy of Digital Instruments.]

consequently has an experimental voltage resolution of a few millivolts. The spatial resolution is determined by the sharpness of the probe and is approximately 50 nm for microfabricated commercial probes.

D. Scanning Capacitance Microscopy (SCM)

Scanning capacitance microscopy (SCM) is a contact mode variant of atomic force microscopy (AFM) in which changes in tip–sample capacitance are imaged simultaneously with surface topography. Samples for SCM analysis are generally covered with a thin insulating layer that serves as the capacitance dielectric between the tip and sample. The spatially mapped capacitance variations can be attributed to changes in the dielectric thickness, changes in the dielectric constant, or variation in the local charge carrier density underneath the dielectric. In the most typical application of SCM, which is the characterization of semiconductor devices, capacitance variations are converted via a theoretical model into a dopant concentration profile in the semiconductor underneath the dielectric.

In typical operation, a conducting tip is scanned in contact at constant applied force over the dielectric layer on the sample, generating a topographic image. While the probe is scanned, a set of electronics (the capacitance sensor) is used to apply an oscillating voltage to the probe and to measure the tip–sample capacitance. The sensor measures dynamic capacitance change rather than static capacitance. A sensor typically consists of a high-frequency oscillator and associated circuitry whose resonant frequency is determined by the total capacitance of the tip–sample assembly. The instrument signal arises from the variation in resonant frequency of the circuit due to tip–sample capacitance changes as the tip scans over the sample. Commercial sensors can detect capacitance changes of 10^{-21} F. The tip and sample must have an insulator between them, and surface preparation is very important to obtaining good SCM images. Resolution is now regularly 10–20 nm.

The most important application of SCM is two-dimensional dopant concentration profiling in semiconductors. However, the raw capacitance–voltage data obtained from SCM must be converted by a mathematical model into a dopant concentration. Therefore, development and validation of appropriate models represents a large part of SCM methodology. To validate the various models, other experimental techniques must be employed to measure and verify the dopant profiles independently. SCM data are usually compared to secondary ion mass spectrometry (SIMS) measurements of dopant concentrations. Spreading resistance profiling (SRP) and computer simulations are also employed to check model validity.

Appropriate calibration of SCM has been achieved in important semiconductor materials, such as silicon and gallium arsenide, and SCM has been used successfully to measure dopant profiles across p–n junctions in these materials with 20-nm resolution. Figure 17 shows SCM images of three different dopant regions in a silicon wafer covered with a thin layer of silicon oxide (SiO_2). The contrast in the image corresponds to different concentrations of n-type (regions I and III) and p-type (region II) dopants.

E. Scanning Electrochemical Microscopy (SECM)

In SECM, a tiny electrode is scanned over a surface immersed in electrolyte solution. The electrode probe senses either electrochemical current or potential, and can be used to map chemical reactivity or the distribution of ions across surfaces. A typical amperometric (current sensing) probe, shown in Fig. 18A, is fabricated by sealing a Pt or Au wire or a carbon fiber in a glass capillary. Potentiometric probes are usually ion-selective electrodes that are sensitive to a given ion, for example, Na^+ or Ca^{2+}. The potential measured by these probes is proportional to ion concentration.

There are many possible scanning modes in SECM; three of the most common are shown in Fig. 18B. In feedback mode, an amperometric probe scans the sample under an electrolyte solution containing a redox mediator, i.e., a soluble molecule that can exist in two states of charge (R, reduced, and O, oxidized). A potential is applied to the probe such that mediator molecules that contact the probe are electrochemically oxidized,

$$R \Rightarrow O + e^-,$$

generating a tip current I_{tip}. The current I_{tip} is a function of how close the tip is to the sample surface. For example,

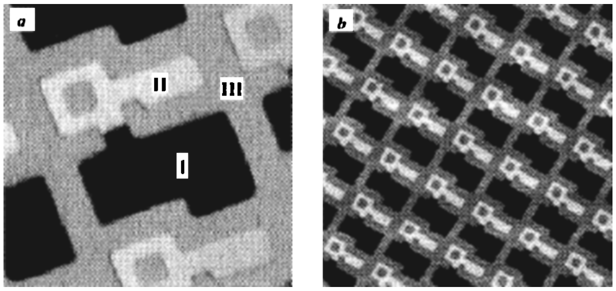


FIGURE 17 Scanning capacitance images of a silicon wafer with an SiO_2 overlayer. Three different dopant regions (contrasts) are visible. Region I: n-type; region II: p-type; region III: highly doped n-type. The images were taken with a 80-kHz, 4-V modulation voltage. The left image is 20 μm × 20 μm. The right image is 80 μm × 80 μm. [Reprinted with permission from Nakagiri, N., Yamamoto, T., Sugimura, H., Suzuki, Y., Miyashita, M., and Watanabe, S. (1997). *Nanotechnology* **8,** A32–A37.]

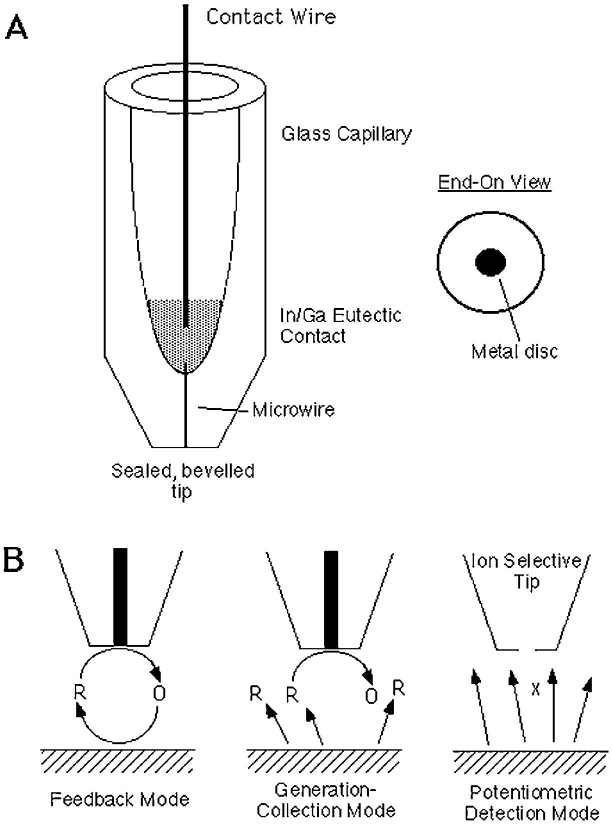


FIGURE 18 (A) Illustration of an SECM amperometric (current-sensing) probe. (B) SECM operation modes.

if the oxidized species O is reduced back to R at the sample,

$$O + e^{-} \Rightarrow R,$$

I_{tip} is increased because of the increased diffusive flux of reduced mediator molecules to the tip. Like STM, I_{tip} can be used to regulate tip–sample separation. However, in contrast to STM, I_{tip} is generated not by electron tunneling, but by electrochemical oxidation and reduction processes coupled with diffusion. Scanning the probe using Z-position feedback to maintain constant I_{tip} generates a topographic image. Perhaps more importantly, the constant-I_{tip} mode can also be used to image chemical heterogeneity of surfaces. For example, if the backreduction of O to R is suppressed on the sample by adventitious contaminants, these contaminated regions will appear as topographic contrast in the SECM image. As in many SPM techniques, interpretation of SECM images does require some independent knowledge of the sample topographic or chemical characteristics.

In some cases the surface chemistry one is interested in following involves molecules that do not undergo reversible electron oxidations and reductions. For these situations, the feedback mode cannot be applied. Instead, a generation–collection approach (Fig. 18B) can be used. In this mode, molecules generated by surface reactions on the sample are collected by the tip and detected amperometrically, i.e., the species is irreversibly oxidized or reduced. There is no feedback mechanism to regulate tip–sample separation in this mode, and consequently the sample must be very flat or the tip must be sufficiently withdrawn to avoid a tip crash. Maps of tip current as a function of position reveal the "hot spots" for chemical reactivity on the surface. For this reason, generation–collection mode SECM can be used to examine the reactivity of solid catalysts in solution. Interestingly, generation–collection SECM has also been applied to the study of chemistry at the interfaces between two immiscible liquids.

Instead of detecting them amperometrically, chemical species diffusing from the surface in generation–collection experiments can be detected potentiometrically with an ion-selective electrode (Fig. 18B). Potentiometric SECM has been used successfully to image dissolution across ionic crystal surfaces, such as Na^{+} dissolution from NaCl surfaces. Potentiometric detection has the advantage that the concentration of diffusing species is not perturbed, i.e., molecules or ions are not electrochemically oxidized or reduced to some other charge state.

In general, SECM resolution is limited by diffusion as well as probe size. Spatial resolution of 200 nm has been demonstrated in favorable cases.

F. Scanning Thermal Microscopy (SThM)

Scanning thermal microscopy (SThM) is a contact AFM technique that allows spatial mapping of temperature or thermal conductivity across a sample surface in addition to topography. Most thermal probes utilize a temperature-sensitive resistor placed on the end of the tip. These resistor probes can be fabricated from a V-shaped Wollaston wire made of a platinum inner core and outer sheath of silver, in which the silver sheath is etched away at the V-shaped tip. Figure 19 shows a Wollaston wire probe. In passive mode, the tip is scanned across a heated sample under constant-force feedback (contact mode) and a small current is passed through the probe to sense the tip resistance. The resistance value at any point is a measure of the local temperature, and thus a temperature map and topographic image may be produced simultaneously.

In active mode, a large current is used to induce Joule heating in the tip resistor. While scanning, heat is conducted away from the tip into the sample, which tends to cool the resistor and change its resistance. A feedback mechanism is employed to keep the current constant. Plotting the voltage required to maintain constant current

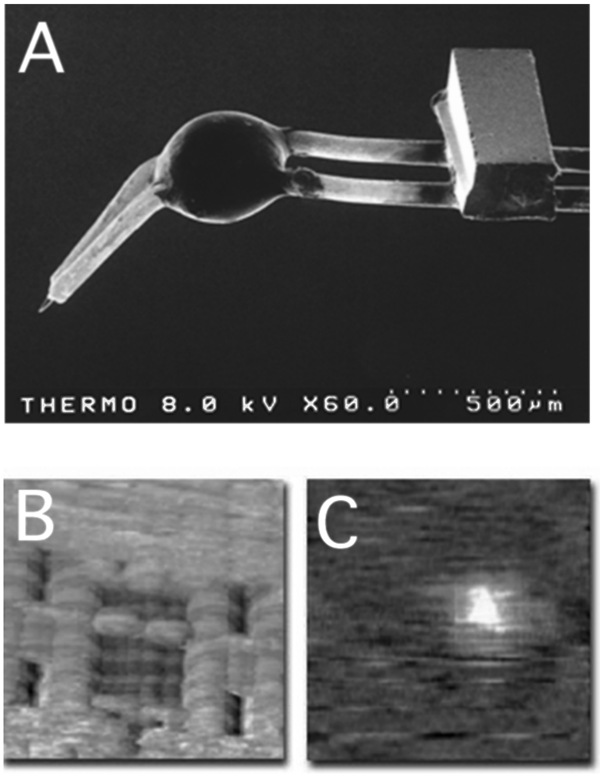


FIGURE 19 (A) Electron micrograph of a Wollaston wire SThM probe. (B, C) Topographic and temperature maps of an active integrated circuit. The bright region in the temperature map (C) reveals a hot spot due to flow of electrical current. [Courtesy of Thermo Microscopes, Inc.]

through the probe as a function of position provides an image of thermal conductance of the sample.

Another recent probe design utilizes a thermocouple junction fabricated on the end of a standard microfabricated AFM tip. This probe is also scanned over a surface using normal force feedback. The temperature of the surface is mapped simultaneously by recording the calibrated voltage between the two leads of the thermocouple.

Temperature mapping can be used to examine electronic devices that are passing current and dissipating energy. Figure 19 shows topographic and temperature maps of a silicon device recorded by SThM using a Wollaston wire probe. SThM has also been applied to a variety of polymer thin-film samples. Typical temperature resolution is tens of millikelvins. Spatial resolution is naturally a function of the probe sharpness and is about 100 nm.

V. FUTURE TRENDS

The resolution and sensitivity of SPM techniques are inherently dependent on the properties of the probe, and efforts are continuous to develop new probes and improve probe fabrication processes. The primary literature is full of new scanning probe concepts, some of which may develop into practical methodologies in the future. NSOM, for example, stands to benefit greatly from the development of low-cost, high-throughput, microfabricated photonic probes as an alternative to the current tapered fiber structures now in use.

SPM is viewed increasingly as a measurement and analysis tool, not just an imaging tool, and consequently efforts are being expended to increase the range of physical properties that can be probed by SPM. One area of significant interest is the development of SPM methods for chemical analysis on sub-100-nm length scales. Nanoscale chemical analysis of surfaces is an important unsolved problem in many areas of industry and science. To be sure, many SPM techniques already provide some type of chemical analysis. For example, STM can qualitatively distinguish between different atoms and chemical functional groups, as shown in Fig. 6. SECM and NSOM provide chemical information by electrochemical current (or potential) and fluorescence, respectively. Still, there is a perceived need to enhance the chemical analysis capabilities of SPM technology by making use of other quantitative and qualitative methods used in analytical chemistry. Efforts to exploit infrared absorption and nuclear magnetic resonance in scanning probe experiments are underway. SPM technology is a very general strategy to imaging and analysis of surfaces, making it likely that new SPM technologies will emerge in the future.

SEE ALSO THE FOLLOWING ARTICLES

Electron Microprobe Analysis of Minerals • Microscopy • Scanning Electron Microscopy • Surface Chemistry • Positron Microscopy • Transmission Electron Microscopy • X-Ray Analysis • X-Ray Photoelectron Spectroscopy

BIBLIOGRAPHY

Barbara, P. F., Adams, D. M., and O'Connor, D. B. (1999). "Characterization of organic thin film materials with near-field scanning optical microscopy (NSOM)," *Ann. Rev. Materials Sci.* **29,** 433–469.

Binnig, G., and Rohrer, H. (1987). "Scanning tunneling microscopy—From birth to adolescence (Nobel Lecture)," *Angew. Chem. Int. Ed. English* **26,** 606–614.

Binnig, G., Quate, C. F., and Gerber, C. (1986). "Atomic force microscope," *Phys. Rev. Lett*. **56,** 930–933.

Carpick, R. W., and Salmeron, M. (1997). "Scratching the surface: Fundamental investigations of tribology with atomic force microscopy," *Chem. Rev*. **97,** 1163–1194.

Dunn, R. C. (1999). “Near-field scanning optical microscopy,” *Chem. Rev*. **99,** 2891–2927.

Hamers, R. J. (1996). “Scanned probe microscopies in chemistry,” *J. Phys. Chem.* **100,** 13103–13120.

Hartmann, U. (1999). “Magnetic force microscopy,” *Annu. Rev. Materials Sci*. **29,** 53–87.

Jacobs, H. O., and Stemmer, A. (1999). “Measuring and modifying the electric surface potential distribution on a nanometre scale: A powerful tool in science and technology,” *Surface Interface Anal*. **27,** 361–367.

Majumdar, A. (1999). “Scanning thermal microscopy,” *Annu. Rev. Materials Sci*. **29,** 505–585.

Mirkin, M. V., and Horrocks, B. R. (2000). “Electroanalytical measurements using the scanning electrochemical microscope,” *Anal. Chim. Acta* **406,** 119–146.

Paesler, M. A., and Moyer, P. J. (1996). “Near-Field Optics: Theory, Instrumentation, and Applications,” Wiley, New York.

Takano, H., Kenseth, J. R., Wong, S.-S., O’Brien, J. C., and Porter, M. D. (1999). “Chemical and biochemical analysis using scanning force microscopy,” *Chem. Rev*. **99,** 2845–2890.

Wiesendanger, R. (1994). “Scanning Probe Microscopy and Spectroscopy,” Cambridge University Press, Cambridge.

Williams, C. C. (1999). “Two-dimensional dopant profiling by scanning capacitance microscopy,” *Annu. Rev. Materials Sci*. **29,** 471–504.

Sonoluminescence and Sonochemistry

Kenneth S. Suslick
University of Illinois

I. Introduction to Cavitation
II. Sonoluminescence
III. Sonochemistry
IV. Summary

GLOSSARY

Cavitation The formation, growth, and collapse of gas and vapor-filled bubbles in a liquid. Irradiation of liquids with sound or ultrasound can create acoustic cavitation; turbulent flow of liquids can create hydrodynamic cavitation.

Heterogeneous sonochemistry The use of high-intensity sound or ultrasound to alter chemical reactions in a two-phase system, usually a liquid and a solid.

Homogeneous sonochemistry The use of high-intensity sound or ultrasound to alter chemical reactions in a single liquid.

Multi-bubble sonoluminescence (MBSL) Emission of light from a cloud of cavitating bubbles formed during ultrasonic irradiation of a liquid.

Single-bubble sonoluminescence (SBSL) Emission of light from a single cavitating bubble in a liquid, usually water.

Sonochemistry The use of high-intensity sound or ultrasound to alter chemical reactions.

Sonoluminescence Emission of light during ultrasonic irradiation of liquids.

SURPRISINGLY, when liquids are exposed to intense ultrasound, high-energy chemical reactions occur, often accompanied by the emission of light. Acoustic cavitation is responsible for both sonochemistry and sonoluminescence. Bubble collapse in liquids results in an enormous concentration of energy from the conversion of the kinetic energy of liquid motion into heating of the contents of the bubble. The high local temperatures and pressures, combined with extraordinarily rapid cooling, provide a unique means for driving chemical reactions under extreme conditions. There are three classes of sonochemical reactions: so-called homogeneous sonochemistry of liquids, heterogeneous sonochemistry of liquid–liquid or liquid–solid systems, and sonocatalysis (which overlaps the first two). Sonoluminescence may generally be considered a special case of homogeneous sonochemistry. In some cases, ultrasonic irradiation can increase reactivity by nearly a millionfold. A diverse set of applications of

ultrasound to enhance chemical reactivity has been explored, with important applications in mixed phase synthesis, materials chemistry, and biomedical uses. For example, the sonochemical decomposition of volatile organometallic precursors in low-volatility solvents produces nanostructured materials in various forms with high catalytic activities. Nanostructured metals, alloys, carbides and sulfides, nanometer colloids, and nanostructured supported-catalysts can all be prepared by this general route. Another important application of sonochemistry to materials chemistry has been the preparation of biomaterials, most notably protein microspheres. Especially for liquid–solid reactions, the rate enhancements that ultrasound can provide have proved extremely useful for the synthesis of organic and organometallic compounds. Because cavitation can only occur in liquids, chemical reactions are not generally seen in the ultrasonic irradiation of solids or solid-gas systems.

I. INTRODUCTION TO CAVITATION

When a liquid is irradiated with high-intensity sound or ultrasound, acoustic cavitation (the formation, growth, and implosive collapse of bubbles in liquids irradiated with sound) generally occurs. This is the phenomena responsible for sonochemistry and sonoluminescence. During cavitation, the collapse of bubbles produces intense local heating and high pressures, with very short lifetimes. In clouds of cavitating bubbles, these hot-spots have equivalent temperatures of roughly 5000 K, pressures of about 1000 atmospheres, and heating and cooling rates above 10^{10} K/s. In single bubble cavitation, conditions may be even more extreme. Cavitation, then, can create extreme physical and chemical conditions in otherwise cold liquids.

If liquids containing solids are irradiated with ultrasound, related phenomena can occur. Near an extended solid surface, cavity collapse becomes nonspherical, which drives high-speed jets of liquid into the solid surface. These jets and associated shock waves can cause substantial surface damage and expose fresh, highly heated surfaces. In addition, high-velocity interparticle collisions will occur during ultrasonic irradiation of liquid–powder suspensions through cavitation and the shockwaves it creates in such slurries. The resultant collisions are capable of inducing dramatic changes in surface morphology, composition, and reactivity.

A. Acoustic Cavitation

Ultrasound spans the frequencies of roughly 15 kHz to 1 GHz. With typical sound velocities in liquids of ≈1500 m/s, acoustic wavelengths range from roughly 10 to 10^{-4} cm. These are not molecular dimensions. Consequently, the chemical effects of ultrasound do not arise from a direct interaction with molecular species: no direct coupling of the acoustic field on a molecular level is responsible for sonochemistry or sonoluminescence. Instead, sonochemistry and sonoluminescence derive principally from acoustic cavitation, which serves as an effective means of concentrating the diffuse energy of sound. Compression of a gas generates heat. When the compression of bubbles occurs during cavitation, it is more rapid than thermal transport and consequently generates a short-lived, localized hot-spot. There is a general consensus that this hot-spot is the source of homogeneous sonochemistry. In 1917, Rayleigh's mathematical model for the collapse of cavities in incompressible liquids predicted enormous local temperatures and pressures. Ten years later, Richards and Loomis reported the first chemical and biological effects of ultrasound.

If a moderately intense acoustic field (greater than ≈0.5 MPa) is applied to a liquid, the liquid can fail during the expansion (i.e., tensile or negative pressure) portion of the sound field; weak sites within the liquid (e.g., preexisting gas pockets, called "cavitation nuclei") are caused to rapidly grow, thereby producing vapor and gas-filled cavities (i.e., bubbles). These bubbles continue to grow during the negative pressure portion of the sound field, until the sound field pressure turns positive. The resulting inertial implosion of the bubbles (now mostly filled with vapor and thus unable to provide stiffness) can be extremely violent, leading to an enormous concentration of energy within the small residual volume of the collapsed bubble (Fig. 1). This violent cavitation event has been termed "transient cavitation." A normal consequence of this unstable growth and subsequent collapse is that the cavitation bubble itself is destroyed. Gas-filled remnants

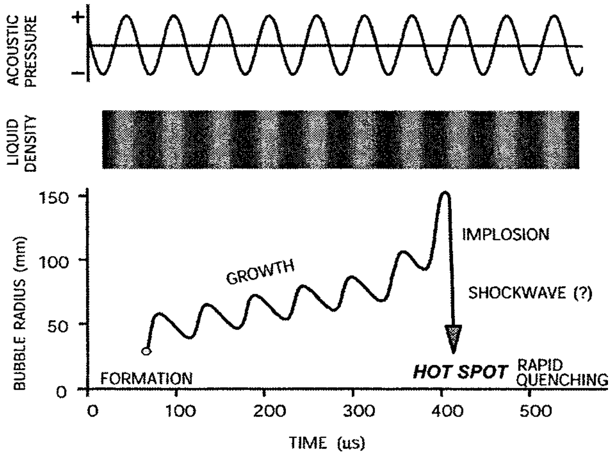


FIGURE 1 Transient acoustic cavitation: the origin of sonochemistry and sonoluminescence.

from the collapse, however, may serve as nucleation sites for subsequent cycles.

For the generally accepted hot-spot theory, the potential energy of the bubble increases as it expands to maximum size, and this energy is then spatially and temporally concentrated into a heated gas core as the bubble implodes. The oscillations of a gas bubble driven by an acoustic field are generally described by "Rayleigh–Plesset" equation; one form of which, called the Gilmore equation, can be expressed a second-order nonlinear differential equation given as

$$R\left(1-\frac{U}{C}\right)\frac{d^2R}{dt^2}+\frac{3}{2}\left(1-\frac{U}{3C}\right)\left(\frac{dR}{dt}\right)^2 - \left(1+\frac{U}{C}\right)H-\frac{R}{C}\left(1-\frac{U}{C}\right)\frac{dH}{dt}-0. \quad (1)$$

The radius and velocity of the bubble wall are given by R and U, respectively. The values for H, the enthalpy at the bubble wall, and C, the local sound speed, may be expressed as follows, using the Tait equation of state for the liquid.

$$H=\frac{n}{n-1}\frac{A^{1/n}}{\rho_0}\left[(P(R)+B)^{n-1/n}-(P_\infty(t)+B)^{n-1/n}\right] \quad (2)$$

and

$$C=\left[c_0^2+(n-1)H\right]. \quad (3)$$

The linear speed of sound in the liquid is c_0. A, B, and n are constants which should be set to the appropriate values (for water these values are $A=3001$ atm., $B=A-1$, and $n=7$). The term $P_\infty(t)$ is the pressure far from the bubble, and includes the ambient pressure plus an appropriate acoustic forcing function. The pressure at the bubble wall (assuming an ideal gas obeying the polytropic law) is given by

$$P(R)=\left(P_0+\frac{2\sigma}{R}\right)\left(\frac{R_0}{R}\right)^{3\gamma}-\frac{2\sigma}{R}-\frac{4\mu U}{R}, \quad (4)$$

where the initial radius of the bubble at time zero is R_0. The ambient pressure of the liquid is P_0, the surface tension σ, the shear viscosity μ, and the polytropic exponent γ.

The validity of the Gilmore equation to compute the behavior of a single, isolated cavitating bubble has been experimentally confirmed. For example, using a light scattering technique, various researchers have obtained measurements of the radius-time curve for single cavitating bubbles (Fig. 2), simultaneous with optical emission from sonoluminescence (see following). The single-bubble sonoluminescent emission is seen as the sharp spike, appearing at the final stages of bubble collapse. Note that these emissions occur at the point of minimum bubble size, and that the general shape of the theoretical radius-time curve is reproduced.

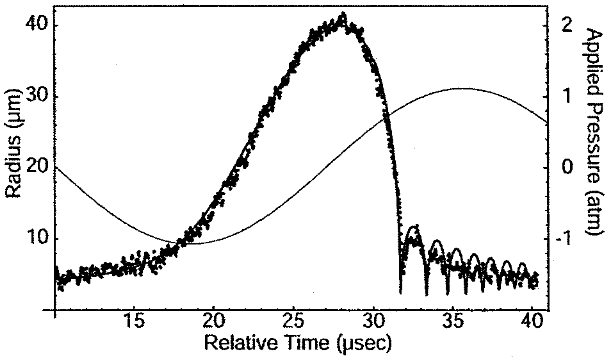


FIGURE 2 Radius–time curves for single cavitating bubbles. A laser is used as a light source to scatter light off the bubble. The scattered light is collected with a lens and focused onto a photomultiplier tube (PMT). The intensity of the scattered light gives the bubble radius, using Mie scattering theory. The collected scattered light is fit to the Gilmore equation, in this case for $R_0 = 5.25$ μm, $P_a = 1.40$ atm, and $R_{max} = 35$ μm, $f = 33.8$ kHz. [From Suslick, K. S., and Matula, T. J. (1999). *Wiley Encyclopedia of Electrical & Electronics Engineering*, Webster, J. G., ed., Wiley-Interscience, New York, vol. 22, p. 648. With permission.]

B. Two-Site Model of Sonochemical Reactivity

Unfortunately, the complex environment that is present in a cavitation field, in which hundreds or thousands of cavitation bubbles interact during their transient cavitation behavior, precludes conventional measurement of the conditions generated during bubble collapse. Chemical reactions themselves, however, can be used to probe reaction conditions. The effective temperature realized by the collapse of clouds of cavitating bubbles can be determined by the use of competing unimolecular reactions whose rate dependencies on temperature have already been measured. This technique of "comparative-rate chemical thermometry" was used by Suslick, Hammerton, and Cline to first determine the effective temperature reached during cavity collapse. The sonochemical ligand substitutions of volatile metal carbonyls were used as these comparative rate probes. These kinetic studies revealed that there were in fact *two* sonochemical reaction sites: the first (and dominant site) is the bubble's interior gas phase while the second is an *initially* liquid phase. The latter corresponds either to heating of a shell of liquid around the collapsing bubble or to droplets of liquid ejected into the hot-spot by surface wave distortions of the collapsing bubble, as shown schematically in Fig. 3.

In addition, for both sites an effective local temperatures was determined by combining the relative sonochemical reaction rates with the known temperature behavior of these reactions. The effective temperature of

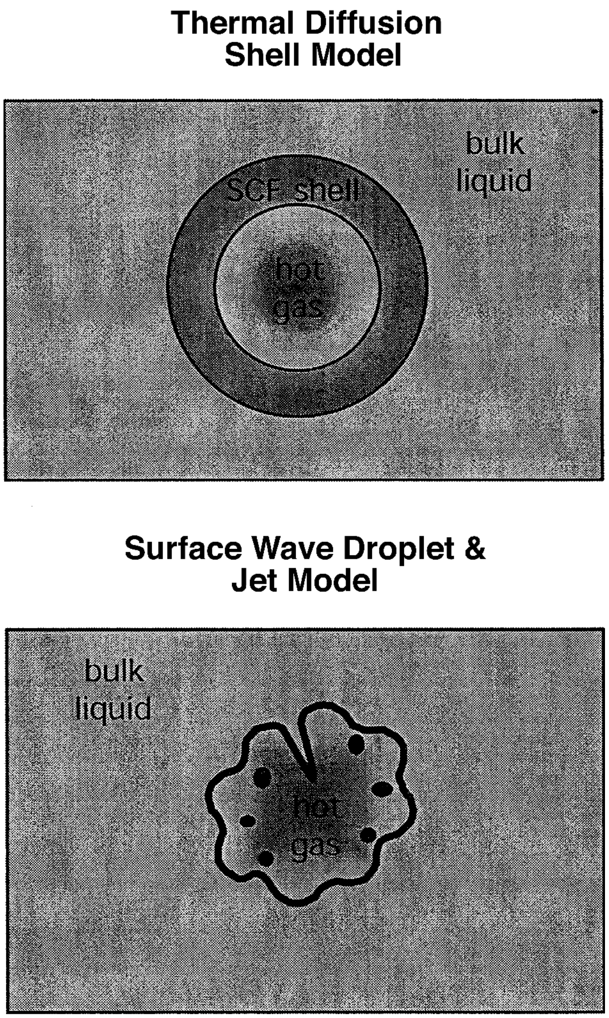


FIGURE 3 Two-site models of the sonochemical reactions sites.

these hot-spots was measured at ≈5200 K in the gas-phase reaction zone and ≈1900 K in the initially liquid zone. Of course, the comparative rate data represent only a composite temperature: during the collapse, the temperature has a highly dynamic profile, as well as a spatial temperature gradient. This two-site model has been confirmed with other reactions and alternative measurements of local temperatures by multi-bubble sonoluminescence are consistent, as discussed later.

C. Microjet Formation during Cavitation at Liquid–Solid Interfaces

Very different phenomena occur for cavitation near extended liquid–solid interfaces. There are two mechanisms for the effects of cavitation near surfaces: microjet impact and shockwave damage. Whenever a cavitation bubble is

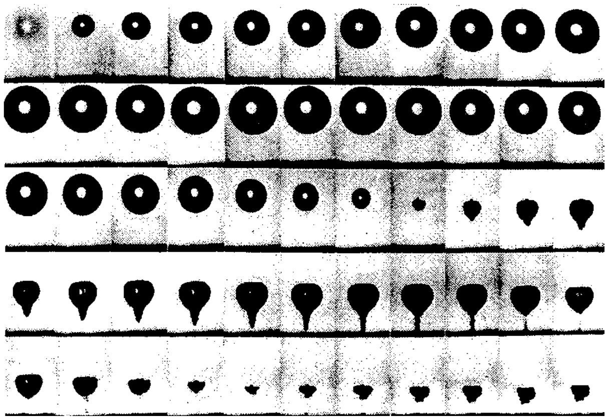


FIGURE 4 Photograph of liquid jet produced during collapse of a cavitation bubble near a solid surface. The width of the bubble is about 1 mm. [From Lauterborn, W., and Vogel, A. (1984). *Annu. Rev. Fluid Mech.*, **16,** 223. With permission.]

produced near a boundary, the asymmetry of the liquid particle motion during cavity collapse induces a deformation in the cavity. The potential energy of the expanded bubble is converted into kinetic energy of a liquid jet that extends through the bubble's interior and penetrates the opposite bubble wall. Because most of the available energy is transferred to the accelerating jet, rather than the bubble wall itself, this jet can reach velocities of hundreds of meters per second. Because of the induced asymmetry, the jet often impacts the local boundary and can deposit enormous energy densities at the site of impact, especially for larger bubbles (i.e., lower frequency). Figure 4 shows a photograph of a jet developed in a collapsing cavity. The second mechanism of cavitation-induced surface damage invokes shockwaves created by cavity collapse in the liquid. The impingement of microjets and shockwaves on the surface create the localized erosion responsible for ultrasonic cleaning and many of the sonochemical effects on heterogeneous reactions. The erosion of metals by cavitation generates newly exposed, highly heated surfaces. Such energy concentration can result in severe damage to the boundary surface; this is less true at higher (MHz) frequencies, simply because the cavitation bubbles are much smaller. This explains the increasing interest in high-frequency ultrasonic cleaning for microelectronics (which has been given unfortunate marketing nom-de-guerre of "megasonics").

In order to induce substantial distortions during bubble collapse, the solid surface must be several times larger than the resonance bubble size: at ≈20 kHz, jet formation becomes important if the solid particles are larger than ≈200 μm. For smaller particles, the shockwaves created by homogeneous cavitation can create high-velocity interparticle collisions, as discussed later.

II. SONOLUMINESCENCE

A. Types of Sonoluminescence

Ultrasonic irradiation of liquids can also produce light, termed "sonoluminescence," as first observed from water in 1934 by Frenzel and Schultes. As with sonochemistry, sonoluminescence derives from acoustic cavitation. There are two classes of sonoluminescence: multiple-bubble sonoluminescence (MBSL) and single-bubble sonoluminescence (SBSL). Since cavitation is a nucleated process and liquids generally contain large numbers particulates that serve as nuclei, the "cavitation field" generated by a propagating or standing acoustic wave typically consists of very large numbers of interacting bubbles, distributed over an extended region of the liquid. Such cavitation can be sufficiently intense to produce MBSL.

For rather specialized but easily obtainable conditions, it is now established that a single, stable gas bubble can be forced into such large amplitude pulsations that it produces sonoluminescence emissions on each (and every) acoustic cycle. This phenomenon is called single-bubble sonoluminescence (SBSL). Under the appropriate conditions, the acoustic force on a bubble can be used to balance against its buoyancy, holding the bubble stable in the liquid by acoustic levitation. This permits examination of the dynamic characteristics of a single cavitating bubble in considerable detail, from both a theoretical and an experimental perspective. Such a bubble is quite small, compared to an acoustic wavelength (e.g., at 20 kHz, the maximum bubble size before collapse is ~50 μm and at minimum during collapse <1 μm).

B. Multiple-Bubble Sonoluminescence

The sonoluminescence of aqueous solutions has been studied extensively over the past 30 years. The spectrum of MBSL in water consists of a peak at 310 nm and a broad continuum throughout the visible region. An intensive study of aqueous MBSL was conducted by Verrall and Sehgal and later by Didenko. The emission at 310 nm is from excited-state $OH^\bullet$, but the continuum is difficult to interpret. MBSL from aqueous and alcohol solutions of many metal salts have been reported and are characterized by emission from metal atom excited states.

Flint and Suslick reported the first MBSL spectra of organic liquids. With various hydrocarbons, the observed emission is from excited states of C_2 ($d^3\Pi_g - a^3\Pi_u$, the Swan lines), the same emission seen in flames. Furthermore, the ultrasonic irradiation of alkanes in the presence of N_2 (or NH_3 or amines) gives emission from CN excited states, but not from N_2 excited states. Emission from N_2 excited states would have been expected if the MBSL originated from microdischarge, whereas CN emission is typically observed from thermal sources. When oxygen is present, emission from excited states of CO_2, $CH^\bullet$, and $OH^\bullet$ is observed, again similar to flame emission.

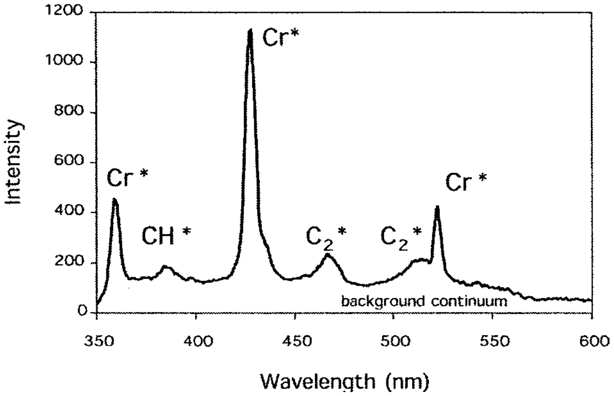


FIGURE 5 Typical MBSL spectrum from a metal carbonyl solution in silicone oil. [From Suslick, K. S. *et al.* (1999). *Phil. Trans. Roy. Soc. London A*, **357,** 340. With permission.]

Ultrasonic irradiation of volatile organometallics (such as $Fe(CO)_5$ or $Cr(CO)_6$) in a low volatility organic liquid produces intense sonoluminescence that corresponds to the known atomic emission lines of the metals, again analogous to flame emission. Hot-spot temperatures are sufficient not only to dissociate all the CO ligands from the metal complex, but also to produce excited state metal atoms. Figure 5 shows a typical MBSL spectrum from a metal carbonyl solution ($Cr(CO)_6$ in this example). Note the intense line emission from the metal atom excited states as well as bands from excited states of the diatomics, C_2 and CH. This metal atom emission provides a useful spectroscopic thermometer, as described later.

For both aqueous and nonaqueous liquids, the emission spectra from MBSL suggests that the principal source of light emission is from chemical reactions involving high-energy species formed during cavitation by bubble collapse. MBSL is principally a form of chemiluminescence, just as flame emission is.

C. Single-Bubble Sonoluminescence

At the time of this writing, SBSL remains under active investigation with unsettled controversies. Theoretical interpretations of the experimental findings continue to be refined. It is not yet possible to provide a definitive mechanism for the light emission process, although the most favored model involves compressional heating (though probably without a convergent shockwave) of the bubble contents, similar to MBSL. The spectra of MBSL and SBSL, however, show some dramatic differences. While MBSL is generally dominated by atomic and molecular emission lines, SBSL is an essentially featureless

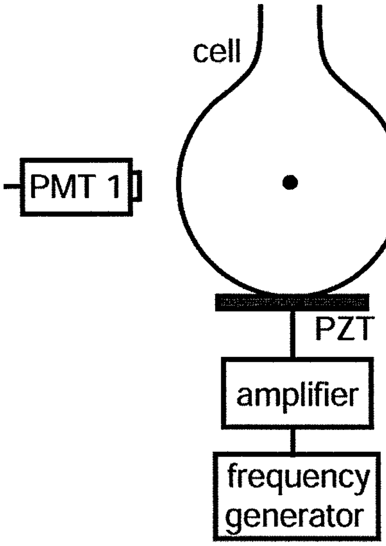


FIGURE 6 A single-bubble sonoluminescence apparatus. A piezoelectric transducer (PZT), mounted to a water-filled levitation cell, is driven by a frequency generator/power amplifier combination. The frequency generator drives the PZT at the appropriate frequency to generate a standing acoustic wave profile within the levitation cell. For experiments to measure the pulse duration from single-bubble sonoluminescence, a time-correlated single-photon counting method works so long as the bubble remains stable and light emission occurs in a synchronous fashion.

emission that increases with decreasing wavelength. For example, an aqueous solution of NaCl shows evidence of excited states of both $OH^{\bullet}$ and Na in the MBSL spectrum; however, the SBSL spectrum of an identical solution shows no evidence of either of these peaks. Nevertheless, the commonality of cause (acoustic cavitation) and effect (light emission) suggests some association in the underlying physics of sonoluminescence for both MBSL and SBSL.

Figure 6 illustrates a typical experimental setup for generating SBSL. A piezoelectric, mounted to a water-filled acoustic levitation cell, is driven to set up a standing wave within the water. The drive frequency depends on the size and geometry of the levitation cell (which can be spherical, cylindrical, or even rectangular). The water is typically degassed to about 10% of saturation. A bubble is introduced by injecting air through a syringe into the water. The large bubbles rise to the surface, while the small bubbles are attracted to pressure antinodes. The final size of the remaining bubble at the antinode depends on gas diffusion steady-state conditions and instabilities present: if the bubble is too small, gas will transport into the bubble; if the bubble is too large, small microbubbles will be ejected from the main bubble. In this manner, the final bubble comes into a diffusive steady state. Once the bubble is positioned at the pressure antinode, the drive pressure amplitude is increased until sonoluminescence is observed.

The radial motion of the bubble was illustrated in Fig. 2. During the main collapse of the bubble, the interior heats up and at the final stages of collapse, light is emitted. With SBSL, the light emission process may occur each and every acoustic cycle, with a synchronicity better than 1 part per billion; for instance, in a 20-kHz sound field (with a period of 50 μs), the light emission can have a jitter of less than 50 ps.

One intriguing aspect of SBSL is the extremely short duration of the sonoluminescence flash. The measured pulse duration of the light flash has been shown to be below 200 ps, and possibly less than 50 ps in some cases. Due to the low levels of light output from a sonoluminescent bubble, one cannot simply use picosecond-response photodiodes. More sophisticated experiments are required. In this case, time-correlated single-photon counting is used to measure the pulse duration. Since the experiment measures the time difference between two photons occurring during the same flash, this measurement produces an autocorrelation of the pulse. Due to the complex and transient nature of cavitation fields, one cannot employ such averaging techniques to MBSL. The most recent studies of MBSL have shown that for aqueous systems involving air and noble gases, the pulse width is also extremely short, less than 1 ns.

The most plausible explanation for the differences between MBSL and SBSL is simply the degree of compression and the extent of consequent local heating. In SBSL, the bubble collapse is much more spherical than is likely in the complex acoustics of a bubble cloud. As a consequence, the effective temperature reached in single-bubble cavitation is probably sufficiently high to induce significant ionization and plasma formation. In these circumstances, SBSL will be dominated by featureless bremstrahlung emission, rather than bands from atomic or molecular emission as in MBSL.

In keeping with this hypothesis, Suslick and coworkers recently found a series of polar aprotic liquids that can support very strong SBSL, and for the first time, observed line emission from molecular excited states during SBSL. This provides a spectroscopic bridge between SBSL and MBSL and gives direct proof of the existence of chemical reactions and the formation of molecular excited states during single-bubble cavitation. In these liquids, both stationary and moving single bubbles could be created. As the sphericity of bubble collapse increases (from MBSL to moving bubble SBSL to stationary bubble SBSL), the efficacy of compression increases, the effective temperatures within the bubble increase, and the emission changes from spectra dominated by excited state molecular emission to featureless (Bremsstrahlung-like) spectra (Fig. 7).

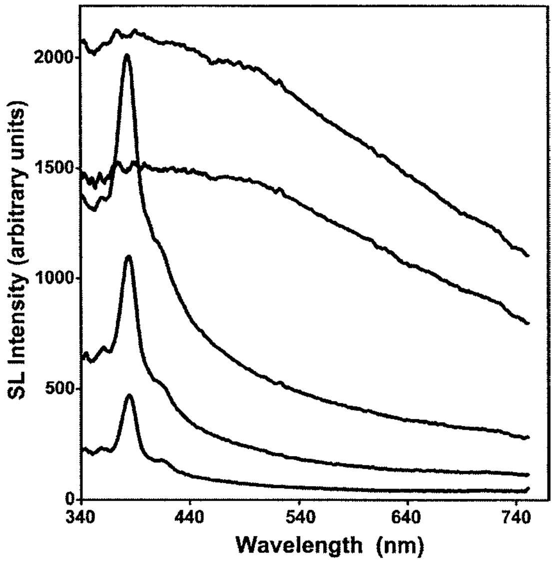


FIGURE 7 Moving, single-bubble sonoluminescence spectra of adiponitrile. Acoustic pressure increases from bottom to top from 1.7 to 1.9 B. The excited state CN comes from the liquid vapor rather than from the nitrogen gas initially present inside the bubble. [From Didenko, Y., McNamara III, W. R., and Suslick, K. S. (2000). *Nature*, **406,** 878. With permission.]

Another major difference between SBSL and MBSL may lie in the contents of the cavitating bubbles. For MBSL, it is generally accepted that a particular bubble in the cavitation field only lasts for a few acoustic cycles before being destroyed, and therefore its contents represent the equilibrium vapor pressures of the solution and its dissolved gases. In contrast, in SBSL, a single bubble can remain stable, emitting light for hours. For air bubbles in water, it is now generally assumed that nitrogen and oxygen molecules dissociate because of the high temperatures, forming NO_x compounds that dissolve in the surrounding water, leaving behind only the nonreactive argon inside the bubble. Thus, even though argon represents only a small fraction of the air concentration dissolved in water, the SBSL bubble acts as a chemical reaction chamber that rectifies argon over thousands of acoustic cycles, until the bubble contents are mostly rarefied argon.

D. Spectroscopic Probes of Cavitation Conditions

Determination of the temperatures reached in cavitating bubbles has remained a difficult experimental problem. As a spectroscopic probe of the cavitation event, MBSL provides a solution. High-resolution MBSL spectra from silicone oil under Ar have been reported and analyzed. The observed emission comes from excited states of diatomic carbon (C_2) and has been modeled with synthetic spectra as a function of rotational and vibrational temperatures. From comparison of synthetic to observed spectra, the effective cavitation temperature is 5050 ± 150 K. The excellence of the match between the observed MBSL and the synthetic spectra provides definitive proof that the sonoluminescence event is a thermal, chemiluminescence process. The agreement between this spectroscopic determination of the cavitation temperature and that made by comparative rate thermometry of sonochemical reactions is surprisingly close.

A second spectroscopic thermometer comes from the relative intensities of atomic emission lines in the sonoluminescence spectra of excited state metal atoms produced by sonolysis of volatile Fe, Cr, and Mo carbonyls. Sufficient spectral information about emissivities of many metal atom excited states are available to readily calculate emission spectra as a function of temperature. Because of this, the emission spectra of metal atoms are extensively used by astronomers to monitor the surface temperature of stars. From comparison of calculated spectra and the observed MBSL spectra from metal carbonyls, another measurement of the cavitation temperature can be obtained. The effective emission temperature during cavitation under argon at 20 kHz is 4900 ± 250 K, as shown in the example given in Fig. 8. Again, agreement with prior comparative rate thermometry and the MBSL emission temperature of C_2^* excited states is excellent.

One may also able to control the temperature within the cavitation bubble simply by changing the bubble contents. Upon addition of gaseous hydrocarbons (methane, ethylene, or propane), the observed emission temperatures from Cr atom excited states systematically decrease: just 3% propane in Ar, for example, reduces the measured emission temperature to 2500 K. As polyatomic molecules are added to the bubble contents, the polytropic ratio of the

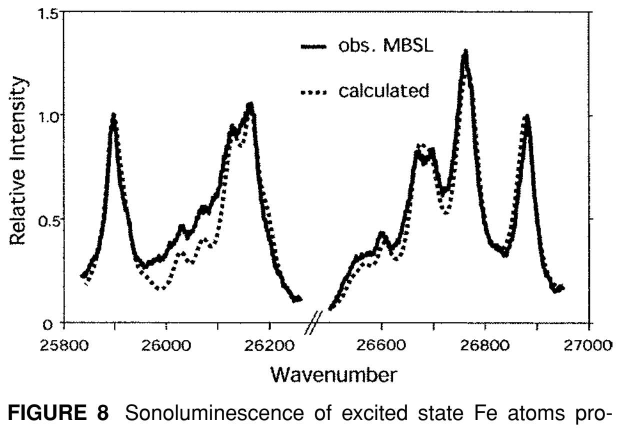


FIGURE 8 Sonoluminescence of excited state Fe atoms produced during sonolysis of $Fe(CO)_5$ dissolved in silicone oil under Ar compared to the calculated spectrum. The emission temperature observed from MBSL from excited state Fe, Cr, and Mo atoms is 4900 ± 250 K. [From Suslick, K. S. *et al.* (1999). *Phil. Trans. Roy. Soc. London A*, **357,** 342. With permission.]

gas in the bubble decreases, and so too does the expected temperature from adiabatic compression. The presence of the polyatomic gas simply provides vibrational and rotational modes that will divert much of the kinetic energy of collapse away from a direct temperature increase. The effects of the addition of polyatomic gases on the observed cavitation emission temperature can be quantitatively modeled by simple adiabatic compression of a bubble during cavitational collapse. This simple model predicts pressures on the order of 10^3 Bar, which is quantitatively consistent with the linewidth broadening and small peak wavelength shifts observed in the metal atom emission. The lifetime of the hot-spot is less well determined. Certainly it is under a microsecond and may be considerably less, given the 200 ps emission lifetimes seen in SBSL. The cooling rates, even at μs lifetimes, are enormous: above 10^{10} K/s.

The interpretation of the spectroscopy of SBSL in water is much less clear. At this writing, SBSL has been observed primarily in aqueous fluids, and the spectra obtained are surprisingly featureless. Some very interesting effects are observed when the gas contents of the bubble are changed. Furthermore, the spectra show practically no evidence of OH emissions, and when He and Ar bubbles are considered, continue to increase in intensity even into the deep ultraviolet. These spectra are reminiscent of black body or bremstrahlung emission with temperatures in excess of 10^4 K. Several other alternative explanations for SBSL have been presented, and there exists considerable theoretical activity in this particular aspect of SBSL.

III. SONOCHEMISTRY

In a fundamental sense, chemistry is the interaction of energy and matter. Chemical reactions require energy in one form or another to proceed: chemistry stops as the temperature approaches absolute zero. One has only limited control, however, over the nature of this interaction. In large part, the properties of a specific energy source determines the course of a chemical reaction. Ultrasonic irradiation differs from traditional energy sources (such as heat, light, or ionizing radiation) in duration, pressure, and energy per molecule. The immense local temperatures and pressures and the extraordinary heating and cooling rates generated by cavitation bubble collapse mean that ultrasound provides an unusual mechanism for generating high energy chemistry. Like photochemistry, very large amounts of energy are introduced in a short period of time, but it is thermal, not electronic, excitation. As in flash pyrolysis, high thermal temperatures are reached, but the duration is very much shorter (by $>10^4$) and the temperatures are even higher (by 5- to 10-fold). Similar to shock-tube chemistry

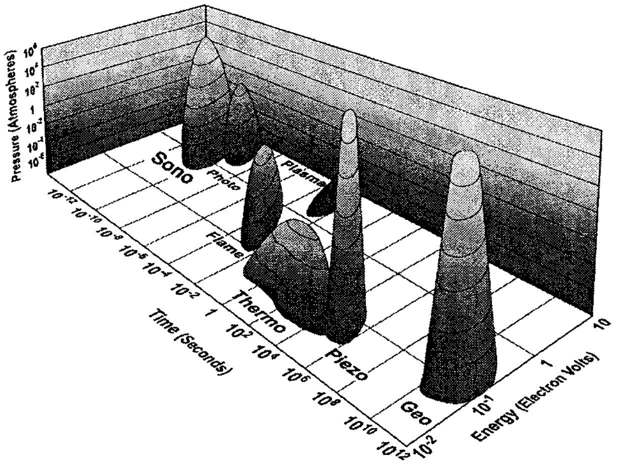


FIGURE 9 Chemistry: the interaction of energy and matter. [From Suslick, K. S. (1990). *Science*, **247,** 1440. With permission.]

or multiphoton infrared laser photolysis, cavitation heating is very short lived, but occurs within condensed phases. Furthermore, sonochemistry has a high-pressure component, which suggests that one might be able to produce on a microscopic scale the same macroscopic conditions of high temperature-pressure "bomb" reactions or explosive shockwave synthesis in solids. Figure 9 presents an interesting comparison of the parameters that control chemical reactivity (time, pressure, and energy) for various forms of chemistry.

The same limitations apply to the control of sonochemical reactions as in any thermal process: the Boltzmann energy distribution means that the energy per individual molecule will vary widely. One does have easy control, however, over the intensity of heating generated by acoustic cavitation using various physical parameters. The origin of these influences is easily understood in terms of the hot-spot mechanism of sonochemistry. The most important parameters are thermal conductivity of dissolved gases (which can effect the degree of adiabaticity), polyatomic content inside the bubble (which reduces compressional heating), and acoustic pressure. As acoustic pressure is increased, there is a threshold value for nucleation and bubble growth and hence sonochemistry, followed by an increase in sonochemical rates due to increased numbers of effectively cavitating bubbles. At sufficiently high intensities, the cavitation of the liquid near the radiating surface becomes so intense as to produce a shroud of bubbles, diminishing penetration of sound into the liquid and decreasing sonochemical rates. In contrast, frequency appears to be less important, at least within the range where cavitation can occur (a few hertz to a few megahertz), although there have been few detailed studies of its role.

Homogeneous sonochemistry typically is not a very energy efficient process, whereas heterogeneous

sonochemistry is several orders of magnitude better. Since ultrasound can be produced with nearly perfect efficiency from electric power, the primary energy inefficiency is due to the small fraction of the acoustic power actually involved in the cavitation events. This might be significantly improved, however, if a more efficient means of coupling the sound field to generate cavitation can be found.

A. Experimental Design

A variety of devices have been used for ultrasonic irradiation of solutions. There are three general designs in use presently: the ultrasonic cleaning bath, the direct immersion ultrasonic horn, and flow reactors. The originating source of the ultrasound is generally a piezoelectric material, usually a lead zirconate titanate ceramic (PZT), which is subjected to a high AC voltage with an ultrasonic frequency (typically 15 to 50 kHz). For industrial use, the more robust magnetostrictive metal alloys (usually of Ni) can be used as the core of a solenoid generating an alternating magnetic field with an ultrasonic frequency. The vibrating source is attached to the wall of a cleaning bath, to an amplifying horn, or to the outer surfaces of a flow-through tube or diaphragm.

The ultrasonic cleaning bath is clearly the most accessible source of laboratory ultrasound and has been used successfully for a variety of liquid-solid heterogeneous sonochemical studies. The low intensity available in these devices ($\approx$1 W/cm^2), however, means that even in the case of heterogeneous sonochemistry, an ultrasonic cleaning bath must be viewed as an apparatus of limited capability. The most intense and reliable source of ultrasound generally used in the chemical laboratory is the direct immersion ultrasonic horn (50 to 500 W/cm^2), as shown in Fig. 10, which can be used for work under either inert or reactive atmospheres or at moderate pressures ($<$10 atmospheres). These devices are available from several manufacturers at modest cost. Commercially available flow-through reaction chambers that will attach to these horns allow the processing of multi-liter volumes. The acoustic intensities are easily and reproducibly variable; the acoustic frequency is well controlled, albeit fixed (typically at 20 kHz). Since power levels are quite high, counter-cooling of the reaction solution is essential to provide temperature control. Large-scale ultrasonic generation in flow-trough configurations is a well-established technology. Liquid processing rates of 200 L/min are routinely accessible from a variety of modular, in-line designs with acoustic power of $\approx$20 kW per unit. The industrial uses of these units include (1) degassing of liquids, (2) dispersion of solids into liquids, (3) emulsification of immiscible liquids, and (4) large-scale cell disruption.

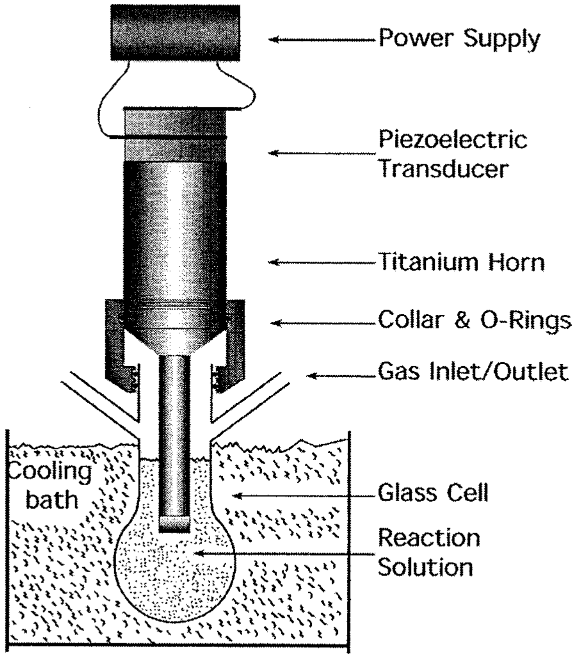


FIGURE 10 A typical sonochemical apparatus with direct immersion ultrasonic horn. Ultrasound can be easily introduced into a chemical reaction with good control of temperature and ambient atmosphere. The usual piezoelectric ceramic is PZT, a lead zirconate titanate ceramic.

B. Homogeneous Sonochemistry: Bond Breaking and Radical Formation

The chemical effect of ultrasound on aqueous solutions have been studied for many years. The primary products are H_2 and H_2O_2; there is strong evidence for various high-energy intermediates, including HO_2, $H^\bullet$, $OH^\bullet$, and perhaps $e^-_{(aq)}$. The work of Riesz and collaborators used electron paramagnetic resonance with chemical spin-traps to demonstrate definitively the generation of $H^\bullet$ and $OH^\bullet$ during ultrasonic irradiation, even with clinical sources of ultrasound. The extensive work in Henglein's laboratory involving aqueous sonochemistry of dissolved gases has established clear analogies to combustion processes. As one would expect, the sonolysis of water, which produces both strong reductants and oxidants, is capable of causing secondary oxidation and reduction reactions, as often observed. Most recently there has been strong interest shown in the use of ultrasound for remediation of low levels of organic contamination of water. The $OH^\bullet$ radicals produced from the sonolysis of water are able to attack essentially all organic compounds (including halocarbons, pesticides, and nitroaromatics) and through a series of reactions oxidize them fully. The desirability of sonolysis for such remediation lies in its low maintenance requirements and the low-energy efficiency of alternative methods (e.g., ozonolysis, UV photolysis).

In contrast, the ultrasonic irradiation of organic liquids has been less studied. Suslick and coworkers established that virtually all organic liquids will generate free radicals upon ultrasonic irradiation, as long as the total vapor pressure is low enough to allow effective bubble collapse. The sonolysis of simple hydrocarbons (for example, n-alkanes) creates the same kinds of products associated with very-high-temperature pyrolysis. Most of these products (H_2, CH_4, and the smaller 1-alkenes) derive from a well-understood radical chain mechanism.

The sonochemistry of solutes dissolved in organic liquids also remains largely unexplored. The sonochemistry of metal carbonyl compounds is an exception. Detailed studies of these systems led to important mechanistic understandings of the nature of sonochemistry. A variety of unusual reactivity patterns have been observed during ultrasonic irradiation, including multiple ligand dissociation, novel metal cluster formation, and the initiation of homogeneous catalysis at low ambient temperature.

C. Applications of Sonochemistry to Materials Synthesis and Catalysis

Of special interest is the recent development of sonochemistry as a synthetic tool for the creation of unusual inorganic materials. As one example, the recent discovery of a simple sonochemical synthesis of amorphous iron helped settle the longstanding controversy over its magnetic properties. More generally, ultrasound has proved extremely useful in the synthesis of a wide range of nanostructured materials, including high surface area transition metals, alloys, carbides, oxides and colloids Sonochemical decomposition of volatile organometallic precursors in high-boiling solvents produces nanostructured materials in various forms with high catalytic activities. Nanometer colloids, nanoporous high surface area aggregates, and nanostructured oxide supported catalysts can all be prepared by this general route, as shown schematically in Fig. 11. There remains much to explore in the sonochemical synthesis of inorganic materials, and this technique has only begun to be exploited.

Heterogeneous catalysis is extremely important in the chemical and petroleum industries, and the applications of ultrasound to catalysis have been reviewed recently. The effects of ultrasound on catalysis can occur in three distinct stages: (i) during the formation of supported catalysts, (ii) activation of pre-formed catalysts, or (iii) enhancement of catalytic behavior during a catalytic reaction. In the cases of modest rate increases, it appears likely that the cause is increased effective surface area; this is especially important in the case of catalysts supported on brittle solids. More impressive accelerations, however, have included hydrogenations and hydrosilations by Ni powder,

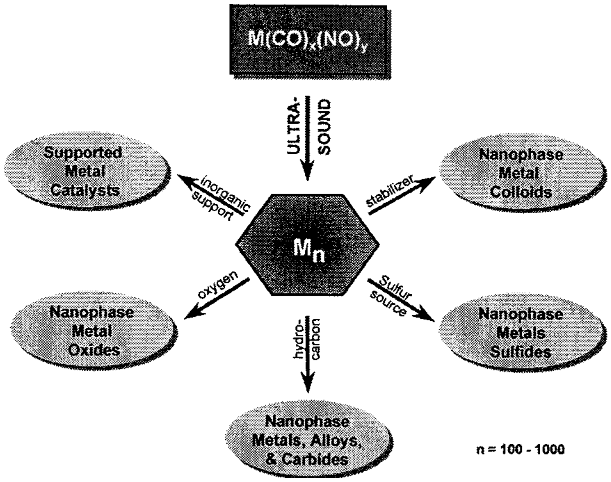


FIGURE 11 Sonochemical synthesis of various forms of nanostructured materials.

Raney Ni, and Pd or Pt on carbon. For example, the hydrogenation of alkenes by Ni powder is enormously enhanced ($>10^5$-fold) by ultrasonic irradiation. As discussed in the next section, this dramatic increase in catalytic activity is due to the formation of uncontaminated metal surfaces from interparticle collisions caused by cavitation-induced shockwaves.

Heterogeneous catalysts often require rare and expensive metals. The use of ultrasound offers some hope of activating less reactive, but also less costly, metals. As one example, ultrasonic irradiation of solutions of $Mo(CO)_6$ produces aggregates of nanometer-sized clusters of face centered cubic molybdenum carbide. The material was extremely porous with a high surface area and consisted of aggregates of $\approx$2-nm-sized particles. The catalytic properties showed the molybdenum carbide generated by ultrasound is an active and highly selective dehydrogenation catalyst comparable to commercial ultrafine platinum powder.

The sonochemical synthesis of nanostructured molybdenum sulfide provides another example of the utility of sonochemistry to the production of active catalysts. MoS_2 is best known as a standard automotive lubricant; its lubricant properties are due to its layered structure. Planes of molybdenum atoms are sandwiched on both faces by planes of sulfur atoms tightly bonded to the Mo. Interactions between the sulfur planes are weak, thus producing lubrication properties similar to graphite. Of greater interest here, however, MoS_2 is also the predominant hydrodesulfurization catalyst heavily used by the petroleum industry to remove sulfur from fossil fuels before combustion. Sonochemistry provides an unusual morphology of MoS_2 by the irradiation of solutions of molybdenum hexacarbonyl, as shown in Fig. 12. Conventional MoS_2 shows a plate-like morphology typical for such layered materials.

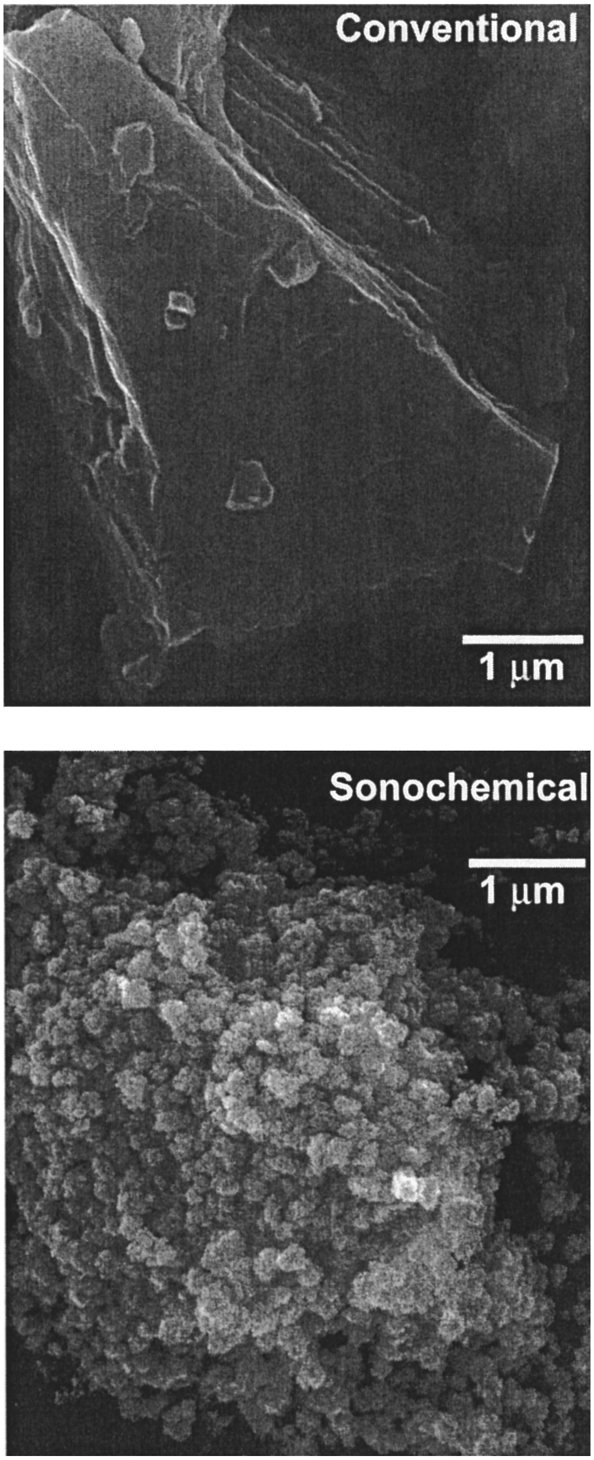


FIGURE 12 Morphology of conventional and sonochemically prepared MoS_2. [From Mdleleni, M. M., Hyeon, T., and Suslick, K. S. (1998). *J. Am. Chem. Soc.*, **120,** 6189–6190. With permission.]

The sonochemical MoS_2 exists as a porous agglomeration of clusters of spherical particles with an average diameter of 15 nm. Despite the morphological difference between the sonochemical and conventional MoS_2, TEM images of both sulfides show lattice fringes with interlayer spacings of 0.62 ± 0.01 nm. The sonochemically prepared MoS_2, however, shows much greater edge and defect content, as the layers must bend, break or otherwise distort to form the outer surface of the 15-nm particle size. It is well established that the activity of MoS_2 is localized at the edges and not on the flat basal planes. Given the inherently higher edge concentrations in nanostructured materials, the catalytic properties of sonochemically prepared shows substantially increased activity for hydrodesulfurization, comparable to those observed with RuS_2, one of the best prior catalysts.

Sonochemistry is also proving to have important applications with polymeric materials. Substantial work has been accomplished in the sonochemical initiation of polymerization and in the modification of polymers after synthesis. The use of sonolysis to create radicals which function as radical initiators has been well explored. Similarly the use of sonochemically prepared radicals and other reactive species to modify the surface properties of polymers is being developed, particularly by G. Price. Other effects of ultrasound on long chain polymers tend to be mechanical cleavage, which produces relatively uniform size distributions of shorter chain lengths.

Another important application has been the sonochemical preparation of biomaterials, most notably protein microspheres. Using high-intensity ultrasound and simple protein solutions, a remarkably easy method to make both air-filled microbubbles and nonaqueous liquid-filled microcapsules has been developed. As shown in Fig. 13, these protein microspheres have a wide range of biomedical applications, including their use as echo contrast agents for sonography, magnetic resonance imaging contrast enhancement, drug delivery, among others, and have generated a substantial patent estate. The microspheres are stable for months, and being slightly smaller than erythrocytes, can be intravenously injected to pass unimpeded through the circulatory system. The mechanism responsible for microsphere formation is a combination of *two* acoustic phenomena: emulsification and cavitation. Ultrasonic emulsification creates the microscopic dispersion of the protein solution necessary to form the proteinaceous microspheres. The long life of these microspheres comes from a sonochemical cross-linking of the protein shell. Protein cysteine residues are oxidized during microsphere formation by sonochemically produced superoxide.

D. Heterogeneous Sonochemistry: Reactions of Solids with Liquids

The use of ultrasound to accelerate chemical reactions in heterogeneous systems has become increasingly widespread. The physical phenomena that are responsible

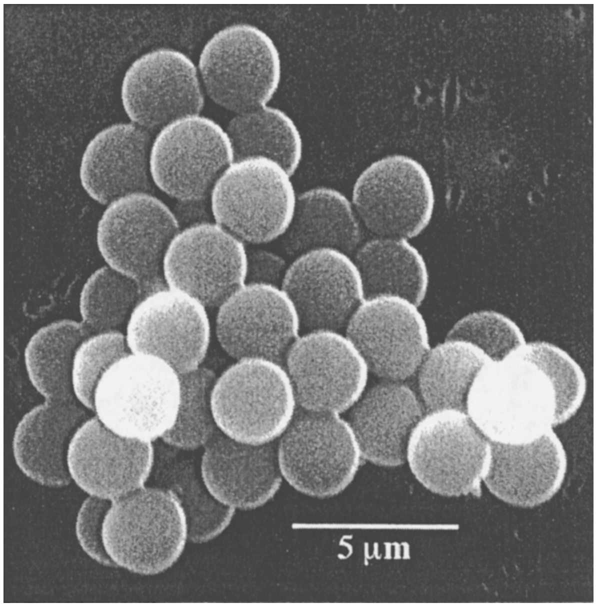


FIGURE 13 Scanning electron micrograph of sonochemically prepared protein microspheres made from hemoglobin. [From Suslick, K. S. (1998). *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th Ed., John Wiley & Sons, New York, vol. 26, p. 534. With permission.]

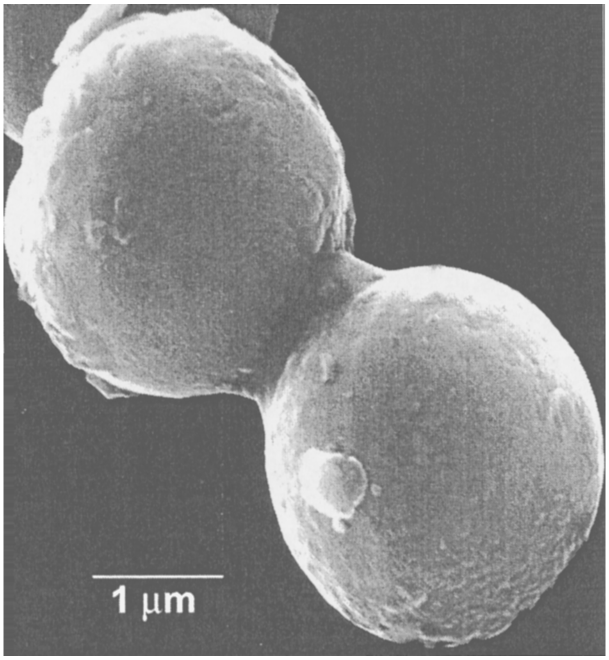


FIGURE 14 Scanning electron micrograph of 5-μm-diameter Zn powder after ultrasonic irradiation of a slurry. Neck formation from localized melting is caused by high-velocity interparticle collisions. Similar micrographs and elemental composition maps (by Auger electron spectroscopy) of mixed metal collisions have also been made. [From Doktycz, S. J. and Suslick, K. S. (1990). *Science*, **247,** 1067. With permission.]

include the creation of emulsions at liquid–liquid interfaces, the generation of cavitational erosion and cleaning at liquid-solid interfaces, the production of shock wave damage and deformation of solid surfaces, the enhancement in surface area from fragmentation of friable solids, and the improvement of mass transport from turbulent mixing and acoustic streaming.

Suslick and coworkers have found that the turbulent flow and shockwaves produced by intense ultrasound can drive metal particles together at sufficiently high speeds to induce effective melting in direct collisions (Fig. 14) and the abrasion of surface crystallites in glancing impacts (Fig. 15). A series of transition metal powders were used to probe the maximum temperatures and speeds reached during interparticle collisions. Using the irradiation of Cr, Mo, and W powders in decane at 20 kHz and 50 W/cm^2, agglomeration and essentially a localized melting occurs for the first two metals, but not the third. On the basis of the melting points of these metals, the effective transient temperature reached at the point of impact during interparticle collisions is roughly 3000°C (which is unrelated to the temperature inside the hot-spot of a collapsing bubble). From the volume of the melted region of impact, the amount of energy generated during collision was determined. From this, a lower estimate of the velocity of impact is roughly one half the speed of sound, in agreement with expected particle velocities from cavitation-induced shockwaves in the liquid.

To enhance the reactivity of reactive metals as stoichiometric reagents, ultrasonic irradiation has become an especially routine synthetic technique for many heterogeneous organic and organometallic reactions particularly those involving reactive metals, such as Mg, Li or Zn. This development originated from the early work of Renaud and the more recent breakthroughs of Luche and others. Examples are shown in Eqs. (5) through (7), where *R* represents an organic functional group and))) represents ultrasonic irradiation. The effects are quite general and apply to reactive inorganic salts and to main group reagents as well. Less work has been done with unreactive metals (e.g., V, Nb, Mo, W), but results here are promising as well. Rate enhancements of more than tenfold are common, yields are often substantially improved, and byproducts avoided.

$$\mathrm{RBr} + \mathrm{R'R''CO} \xrightarrow[\mathrm{Li}]{)))} \mathrm{RR'R''COH} \tag{5}$$

$$\mathrm{RBr} + \mathrm{(CH_3)_2NCHO} \xrightarrow[\mathrm{Li}]{)))} \mathrm{RCHO} \tag{6}$$

$$\mathrm{RR'CO} + \mathrm{BrCH2CO2R''} \xrightarrow[\mathrm{Li}]{)))} \mathrm{RR'C(OH)CH_2CO_2R''}. \tag{7}$$

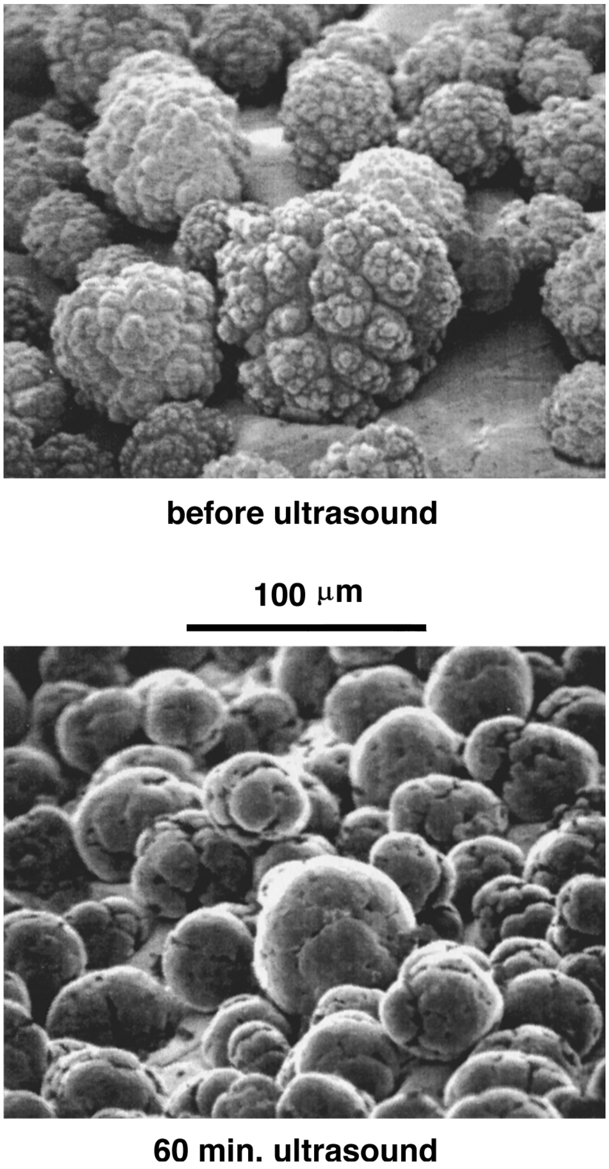


FIGURE 15 The effect of ultrasonic irradiation on the surface morphology and particle size of Ni powder. Initial particle diameters before ultrasound were ≈160 μm; after ultrasound, ≈80 μm. High-velocity interparticle collisions caused by ultrasonic irradiation of slurries are responsible for the smoothing and removal of passivating oxide coating. [From Suslick, K. S. and Doktycz, S. J. (1990). *Advances in Sonochemistry*, Mason, T. J., ed., JAI Press, New York, vol. 1, p. 216. With permission.]

The mechanism of the sonochemical rate enhancements in both stoichiometric and catalytic reactions of metals is associated with dramatic changes in morphology of both large extended surfaces and of powders. As discussed earlier, these changes originate from microjet impact on large surfaces and high-velocity interparticle collisions in slurries. Surface composition studies by Auger electron spectroscopy and sputtered neutral mass spectrometry reveal that ultrasonic irradiation effectively removes surface oxide and other contaminating coatings. The removal of such passivating coatings can dramatically improve reaction rates. The reactivity of clean metal surfaces also appears to be responsible for the greater tendency for heterogeneous sonochemical reactions to involve single electron transfer rather than acid-base chemistry.

Applications of ultrasound to electrochemistry have also seen substantial recent progress. Beneficial effects of ultrasound on electroplating and on organic synthetic applications of organic electrochemistry have been known for quite some time. More recent studies have focused on the underlying physical theory of enhanced mass transport near electrode surfaces. Another important application for sonoelectrochemistry has been developed by J. Reisse and coworkers for the electroreductive synthesis of submicrometer powders of transition metals.

IV. SUMMARY

The phenomenon of acoustic cavitation results in an enormous concentration of energy. The extraordinary local temperatures and pressures so created result in both sonochemistry and sonoluminescence, which provide a unique means for fundamental studies of chemistry and physics under extreme conditions. The chemical consequences of acoustic cavitation are far reaching, both in homogeneous liquids and in mixed-phase system. In the latter, cavitation can have dramatic effects on the reactivities of both extended solid surfaces and on fine powder slurries through microjet and shock wave impact (on large surfaces) and interparticle collisions (with powders). The applications of sonochemistry are diverse and still emerging, especially in the areas of mixed phase synthesis, materials chemistry, and biomedical products.

SEE ALSO THE FOLLOWING ARTICLES

Biomaterials, Synthesis, Fabrication, and Applications • Luminescence • Ultrasonics and Acoustics

BIBLIOGRAPHY

Cheeke, J. D. N. (1997). *Can. J. Phys.* **75,** 77–96.
Crum, L. A. (1994). *Physics Today* **47,** 22.
Crum, L. A., Mason, T. J., Reisse, J., and Suslick, K. S. eds. (1999). "Sonochemistry and Sonoluminescence," Kluwer Publishers, Dordrecht, Netherlands, NATO ASI Series C, v. 524.
Leighton, T. G. (1994). "The Acoustic Bubble," Academic Press, London.

Luche, J.-L., and Bianchi, C. (1998). "Synthetic Organic Chemistry," Kluwer Publishers, Dordrecht, Netherlands.

Mason, T. J. (ed.) (1990, 1991, 1993, 1996, 1999). "Advances in Sonochemistry," vols. 1–5, JAI Press, New York.

Putterman, S. J., and Weninger, K. R. (2000). Sonoluminescence: How bubbles turn sound into light. *Annu. Rev. Fluid Mech*. **32,** 445–476.

Suslick, K. S., and Crum, L. A. (1997). Sonochemistry and sonoluminescence. *In* "Encyclopedia of Acoustics" (M. J. Crocker, ed.), Wiley-Interscience, New York, vol. 1, ch. 26, pp. 271–282.

Suslick, K. S. (1997). Sonocatalysis. *In* "Handbook of Heterogeneous Catalysis" (G. Ertl, H. Knozinger, and J. Weitkamp, eds.), Wiley-VCH, Weinheim, vol. 3, ch. 8.6, pp. 1350–1357.

Suslick, K. S. (1998). Sonochemistry. *In* "Kirk-Othmer Encyclopedia of Chemical Technology," 4th edition, Wiley, New York, vol. 26, pp. 517–541.

Suslick, K. S., Didenko, Y., Fang, M. M., Hyeon, T., Kolbeck, K. J., McNamara III, W. B., Mdleleni, M. M., and Wong, M. (1999). "Acoustic cavitation and its chemical consequences," *Phil. Trans. Roy. Soc. London A* **357,** 335–353.

Suslick, K. S., and Price, G. (1999). "Applications of ultrasound to materials chemistry," *Annu. Rev. Matl. Sci.* **29,** 295–326.

Spectroscopy in Forensic Science

Michael B. Eyring
Micro Forensics Institute, Ltd.

I. History
II. Forensic Spectroscopy Techniques
III. Conclusion

GLOSSARY

Chromatography A term that literally means graphing color or separating colors from a mixture. Today it refers to a wide variety of methods that are used to separate mixtures of compounds so that the separate components can be analyzed and identified.

Dispersive system A component of a spectroscopy system that allows small portions of a spectrum to be separated, identified, and analyzed. A prism disperses white light into its colored components so that they appear as a rainbow. (See monochromator)

Fluorescence Longer wavelength light emitted by a material when it is illuminated by shorter wavelength light in the visible or ultraviolet wavelength regions. Fluorescence can be seen in the bluish glow of quinine in an acidic lime and tonic water drink when it is illuminated by sunlight that contains ultraviolet radiation.

Infrared (IR) radiation Light with wavelengths ranging from about 1000 to 40,000 nm that are commonly perceived as heat. Longer wavelengths of light include microwaves and radio waves.

Monochromator A devise that accepts a broad spectrum of radiation at its input but selectively transmits only a narrow portion or one (mono) color (chroma) of that spectrum at its output. The devise is used to transmit a narrow range of wavelengths of light to a detector or portion of a detector at a time. A monochromator can be a dispersive system.

Nanometer (nm) One thousandth of one millionth of a meter. There are about 60,000 nm in the width of a typical human head hair.

Ultraviolet (UV) radiation Light with a wavelength ranging from about 120 to 360 nm that is commonly associated with sunburns. Shorter wavelengths of light include x-rays and gamma rays.

Visible (VIS) radiation Light with wavelengths ranging from about 380 to 780 nm. The shorter wavelengths appear blue and the longer wavelengths appear red. This wavelength range is referred to as the visible spectrum that is seen in rainbows.

Wavelength The distance from one crest to the next crest of a wave.

SPECTROSCOPY is traditionally the study of radiation that is emitted from, absorbed, or shifted in frequency by its interaction with an object or material. It is generally distinguished from spectrometry, which studies the mass of particles or the energy of radiations that can not generally be dispersed or detected using optical techniques. The distinction between spectroscopy and spectrometry has become somewhat blurred with time and usage but this article will not address topics such as mass spectrometry

(MS), nuclear magnetic resonance (NMR), or energy dispersive x-ray spectrometry (EDX or EDS).

Forensic is a term generally associated with public speaking or formal debate. It refers to speaking before the law or court as it is used in this article and is related to both civil or tort and criminal law. It has become the habit of some to restrict the use of the terms such as forensic science or forensic spectroscopy to criminal cases. In fact, criminal forensic science is one of the smaller divisions within the forensic science community. The majority of forensic science is practiced in other fields. Some of them include occupational health and safety, environmental protection and remediation, personal injury claims resolution, medical malpractice investigations, and environmental toxicology studies, fine art validation and valuation, nuclear safety violation analysis, nuclear materials diversion investigations, and also commercial product failure analysis. These topical areas do not begin to address the rather large areas of social and psychological forensic science and medicine. Any of these fields can find some form of spectroscopy or spectrometry to be of use.

I. HISTORY

Many of the initial methods scientists used to evaluate materials relied on color development or color changes to assist them in the identification of drugs, poisons, explosives, metals, body fluids, plant extracts, and unknown organic and inorganic compounds in general. A detailed summary of such tests was included as a 368-page addendum in the 1940, fifth edition of the Merck Index. Quite a few of those tests were developed in the 1800s and many find use today.

Spectroscopy is defined as the evaluation or study of the production and measurement of electromagnetic radiation spectra, associated with the emission or absorption of radiant energy (light) by different materials. It is a diverse topic and can be as simple as the assessment of the freshness or ripeness of fruit by visual inspection of color. Such an assessment might be the foundation of a produce merchant's insurance claim. More complex visual spectroscopy might include the evaluation of chemical color tests that absorb light or flame color tests of materials that emit light when held in flames on platinum wires. It could also include the colors of materials fused in beads of borax, or the color observations of fine powders formed by scraping a material across an unglazed, white porcelain plate. Both are types of tests which selectively absorb some colors of light and leave other colors to be seen. Some of these tests are still used to evaluate minerals, drugs, metals, and general classes of organic and inorganic materials. In these cases, the eye and brain are acting as the detector and computer that evaluate a spectrum of visible radiation.

Visual spectroscopy can also be semiquantitative. A simple example is the assessment of swimming pool water quality and its chemical content using well-known color reactions and simple sets of colored concentration reference standards. This evaluation of light or color intensity is called photometry. Visual photometry saw one early use in criminal law with the evaluation of breath alcohol concentrations. The test was based on the ability of breath alcohol to chemically reduce the colored chromate ion to a colorless form in aqueous solutions. As more alcohol vapor was bubbled through a chromate solution it became more bleached. The resulting solution was then visually compared to a calibrated, standard set of diluted chromate solutions to determine the subject's breath alcohol content. The test was soon modified to include instrumental measurement of the decolorizing reaction and is referred to as a form of spectrophotometry.

Photographic film was also used as a detector in spectroscopy in order to extend spectral evaluations beyond the limits of the eye. Photographic film is an effective detector but its spectral sensitivity variations, processing, and dimensional instability make it somewhat cumbersome to use. The inclusion of photomultiplier detectors and new radiation sources in the evaluation of spectra opened the way to more easily evaluating materials beyond the visible light region of the radiation spectrum by also using invisible ultraviolet (UV) and infrared (IR) radiation. A very general block diagram of a spectroscopy system is given in Fig. 1.

Visible and UV radiation is of relatively short wavelength and high energy and generally interacts with the electrons around atoms. Infrared radiation is longer wavelength and weaker. It generally effects the bonds between atoms in molecules. Information can be gathered about the nature or environments of those electrons and bonds by evaluating the energy they absorb or emit. The fact that elements, atoms, and molecules can absorb or emit different wavelengths of light or radiation has led scientists to develop a wide array of spectroscopy instruments. Each instrument provides an additional bit of information to the forensic spectroscopist about the material that they may be testing.

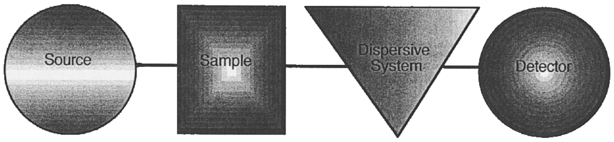


FIGURE 1 Basic components of a spectroscopy system or instrument.

Most spectroscopical analysis is a data intensive process. A single sample scan across a wavelength region of interest can produce a two- or three-dimensional set of data containing 2000 or 3000 data points. A few dozen scans of that sample can yield a data set that was quite cumbersome to handle just a few years ago. It is essential to acknowledge that all forms of spectroscopy and spectrometry have been greatly enhanced by the development of personal computers. When Fourier transform IR spectrometers were first commercialized they filled a room. Today they can fit into a suitcase and be operated in the heart of the wilderness. To be sure, there are trade-offs to be made in instrument performance within these extremes, but computer miniaturization has played a significant part in this development.

The balance of this article will offer a look at a variety of spectroscopical instruments, the materials they can be applied to, and the types of information they can provide.

II. FORENSIC SPECTROSCOPY TECHNIQUES

A. Visible Spectroscopy

Instrumental visible spectroscopy (VISS) is a logical extension of the color tests and visual observations that were mentioned earlier. It is a sensitive, accurate method of measuring the colors or mixtures of colors that our eyes perceive and removes the obvious subjectivity and visual anomalies associated with human vision. It found early and continued use in part because it used simple light sources and its sample containers and optical systems could be made of common glass.

Scientists have known since the early 1900s that they could measure color differences that were too small for the eye to distinguish. Since then, VISS has found wide use in the forensic analysis of materials and systems as diverse as pigments, dyes, and chemical reaction color changes. It is also used in studies of fluorescence, atmospheric contamination, video monitor color correction, toxicological variations in red blood cells, photographic color control, and window film color density measurements, to mention just a few. This versatile technique can be applied to a wide range of sample types from solids and liquids to pastes or gels and the samples can be viewed by either transmitted or reflected light.

VISS is an essential element of color comparisons in criminal cases involving paint and fiber transfers due to its power to distinguish colors the human eye might mistakenly find indistinguishable. Civil investigations also find use of VISS in establishing tampering or trademark violations where the fraudulent material has been painted or dyed to mimic the protected or original product. Different colors or color mixtures that might fool the eye may be easily distinguished with the instrument.

B. Ultraviolet Spectroscopy (UVS)

Ultraviolet spectroscopy requires somewhat more costly quartz glass sample holders and quartz or fluoride optical components as well as a different light source than VISS but has become equally accessible and easy to use. The technique is applied to the evaluation of samples that may contain aromatic compounds such as benzene derivatives or to a wide variety of UV absorbing compounds that are used to protect painted coatings, human skin, and fabrics. Other compounds that are used to whiten fibers and fabrics by producing blue-white fluorescence under outdoor illumination are called fluorescent brightners and may also be analyzed using UVS and UV fluorescence spectroscopy. UVS is also used to detect and quantify compounds that are separated from complex mixtures during liquid chromatography (LC). This method can separate and detect compounds that are easily destroyed by other techniques. LC is also applied to the evaluation of environmental samples for contamination, drug mixtures including steroids, foods, dye mixtures, and complex flavoring agents.

The application of UVS has declined from its earlier use in the criminal analysis of drugs as other more powerful instruments such as gas chromatograph–mass spectrometers (GC–MS) have become less expensive and simpler to operate.

C. Fluorescence Spectroscopy (FS)

Fluorescence spectroscopy instruments illuminate samples with short wavelength light in the visible or ultraviolet wavelength regions. A sample atom's electrons may absorb a portion of the incident light and move to higher energy levels. As that empty energy level is refilled with another electron, the atom may emit a lower energy/longer wavelength light photon that is characteristic of the sample molecule's structure. The emitted light is called fluorescence. Some of the uses of FS were described in the previous section.

FS instrumentation tends to be more complex than UV systems, with monochromators in both illuminating and detector beam paths and, sometimes, dual detectors. These features are necessary to separate rather intense illuminating beams scattered radiation from the sometimes weaker molecular fluorescence.

Fluorescence spectroscopy was once used to identify and quantify difficult to analyze materials like morphine that was extracted from body fluids or lysergic acid diethyl

amide (LSD) recovered from paper or tablet dosage units. It still provides methods for evaluating the effluent of manufacturing processes such as fiber/fabric dyeing, paper production, waste recovery, landfills, and fuel storage facilities.

The instruments tend to be sensitive to temperature, sample acidity and dissolved ion content, agents that quench or suppress fluorescence and need to operated under laboratory conditions. Newer instruments have replaced FS in many of its historical applications but it is reemerging and gaining in use within the forensic science community. This is due to the fact that DNA-based analysis technology is being applied to an ever-increasing group of materials, and fluorescent dyes are used to detect those material's presence in a variety of samples.

It is common to assume that DNA technology is only applied to identifying biological individuals. In fact, that same technology can be, and is being, modified for the analysis of drugs, pollutants, and toxicology samples. Fluorescent compounds are used as tags that make sample components detectable and a wide variety of spectroscopy, photography and photometry systems are used to record the analytical results.

D. Infrared Spectroscopy (IRS)

Infrared spectroscopy is also called molecular spectroscopy because it provides information about a molecule's structure or of a mixture of molecules that are present in the sample. It is one of the most versatile and easily applied forms of spectroscopical analysis. It has been advanced by the development of a host of sampling assessories, detector types, and technological advances that allow its application to all four of the common forms of matter: solids, liquids, gases, and plasmas. Its use in plasma sample analysis in forensic cases is not routine but it is one of the more popular forms of spectroscopy used in forensic science.

1. Infrared Absorption Spectroscopy

Infrared spectroscopy (IRS) is the study or analysis of how the bonds between atoms of a molecule absorb energy and vibrate or bend (deform) much like springs. The energy necessary to cause motions in these bonds is quite specific and provides the forensic scientist with information about how atoms and combinations of atoms are associated in a molecule.

There are some basic requirements for absorption IRS. One is that a vibration or deformation must be associated with a permanent charge separation or electric dipole if it is to be detected. Vibrations between atoms in molecular groups such as the eight-atom sulfur ring in pure sulfur will not be detectable. By comparison, the motion of a carbon–hydrogen bond is generally easily detected. A second limitation lays in the fact that components of a mixture need to be present in excess of about 1–5% if they are to be detectable.

Finally, an IR instrument must be able to "see" the item under test. Waves of radiation are not affected by particles or samples that are much smaller than the wavelength of the radiation. One might note that a beach ball or a sea gull has little affect on an ocean wave and a vacationer standing in shore counting waves would not know if either of those items was floating a hundred yards off shore. When a scientist analyzes radiation that has passed by a sample, he or she can generally not detect ("see") the sample if the illuminating radiation did not interact with it. Much of the important information in an IR spectrum resides in the 10- to 40-μm or 10,000- to 40,000-nm region of the light spectrum. These facts generally limit IRS analysis to samples that are at least 10 μm in size and are mentioned here because they are not usually a factor in other types of spectroscopy.

Infrared spectra can be produced from samples that are imaged by microscopes at distances of less than a millimeter or by telescopes at distances of a kilometer or more. It is used to analyze samples ranging from drug crystals and thin paint layers to atmospheric pollutants.

2. Raman Spectroscopy

Raman spectroscopy (RS) can be considered a second form of infrared analysis that illuminates samples with short wavelength (not infrared) light and then analyzes the light that is scattered at right angles to the illuminating beam. A laser generally provides illumination and molecules in the laser beam absorb light as they vibrate and shift the wavelength of some of the scattered illuminating light by wavelengths comparable to IR absorption bands. These are referred to as Raman shifts or Raman spectra. These spectra are not to be confused with fluorescence spectra that are produced by a different process. This form of IR spectroscopy has the advantage of being able to analyze symmetrical molecules and bonds that do not exhibit permanent charge separations or dipoles and is quite complementary to IRS. It also allows the analysis of Raman shifts and IR spectra that may be outside the detection range of a particular IR detector.

Early Raman spectroscopy instruments saw little forensic use. Illumination sources were mercury vapor lamps with filters that isolated specific mercury radiation wavelengths. Sample sizes were rather large and Raman spectra were captured on photographic film. The detection of weak spectral lines sometimes required days of operation and film exposure. Today's instruments' use of laser

illumination, Fourier transform dispersive systems, new detectors, a variety of sampling devices, and computers have all combined to make Raman spectroscopy a versatile analytical tool.

E. Light Detection and Ranging (LIDAR) Spectroscopy

The fluorescence analysis technique called LIDAR currently requires that specific sample species be identified for analysis. Selected wavelengths of illumination are sent out through a telescope toward a target area and used to excite molecules of the selected compounds. The resulting fluorescence of excited molecules/compounds/species is then collected by a high-performance telescope and detector system. The resulting spectrum is analyzed to determine their concentration of fluorescent species and can be used to evaluate atmospheric pollution levels.

A second form of LIDAR, known as Differential Absorption LIDAR (DIAL), uses two laser beams, one selected to be absorbed by a particular gas or vapor and a second as an unabsorbed reference. Pulses of the two lasers are then aimed and focused at a distant area through a telescope. Gas molecules in that area reflect some of both of the laser beams back to the telescope where they are detected and analyzed to determine if the gas or vapor of interest is present and, if it is, in what concentration.

Areas of pollution, clandestine laboratory operations and the wind drift of pollutants can be mapped by this technique. It can be a particularly powerful tool in forensic investigations because it is a non-intrusive device that can search the atmosphere around a suspected cite of illicit chemical manufacturing, storage, pollution or disposal without requiring direct access to the subject site.

F. Atomic Spectroscopy (AS)

Atomic spectroscopy refers to a wide range of techniques that are applied to materials that have been reduced to atomic or ionic forms or species in an electric arc, flame, or plasma. When one speaks of atomic spectra, molecules and compounds that have been of discussion up to now are no longer an issue. AS is in fact a monumental extension and improvement upon the flame color tests that were mentioned at the beginning of this article.

Molecules and compounds that have been irradiated with a burst of sufficient energy will decompose into their constituent atoms. This energy can also cause the electrons around the atoms to move into higher energy states with a resulting absorption and later emission of specific wavelengths of light. Atomic spectroscopy uses these exchanges of energy to identify atomic species and to allow an investigator to characterize the materials that may have contributed to their presence. The primary forensic spectroscopy advantage of AS lays in its improved sensitivity over molecular spectroscopy and its ability to detect very low concentrations of elements in a complex mixture such as soil or paint.

This sensitivity has led to AS use for the detection and identification of minor elements in samples of glass and to the confirmation or exclusion of tracer elements. Tracer elements are added to materials such as polymers, pigments, or glasses and serve as atomic trade marks for a specific manufacturer's product. Tracers are typically low concentrations of rare earth elements. These tests can be essential in helping forensic scientists discriminate otherwise similar materials in product based lawsuits where a manufacturer claims that he or she did not make the product in question. A few basic techniques of AS are in use and each can have a few variations.

1. Atomic Emission Spectroscopy (AES)

Early AES instruments were rather straightforward and excited the atoms in a sample by placing the sample in a carbon arc. The resulting light was passed through a diffracting prism, or across a diffraction grating, and the resulting light spectrum was recorded on a strip of photographic film. This film was developed and placed in a projection system where the recorded spectral lines were compared with a master film to identify the elements present in the sample.

These systems were effective and some are still in use. Their major drawbacks lay in some tedium in sample preparation, film development and spectral analysis. The recording film also limited the dynamic range of the system (the range of concentrations that could be detected in a single sample) and trace elemental analysis was difficult. These early systems have largely been replaced by a variety of more sensitive, wider range, computer-based equipment. All of the newer equipment is also much more expensive to purchase and operate and requires careful handling and preparation of samples to avoid contamination.

2. Inductively Coupled Plasma-Optical Emission Spectroscopy (ICP-OES)

ICP–OES (or ICP-Atomic Emission Spectroscopy, ICP-AES) systems excite atoms to emit analytically useful radiation using plasmas with temperatures of about 1000°C. Samples are first fused, dissolved in a solvent and then aspirated into the plasma flame. The element specific light emitted by the plasma is then viewed either radially (through the side of the plasma) or axially (through the

length of the plasma) plume and viewed or analyzed by a multi-element spectrometer and detector.

The technique is similar to atomic absorption spectroscopy in sensitivity but its ability to detect several elements at a time improves its utility in forensic analysis. A forensic scientist seldom knows what possible elements that might be found in a sample and multiple element detection is an asset. ICP has been made even more sensitive, especially to metals, by coupling the plasma plume or flame to a mass spectrometer (ICP-MS).

3. Flame Atomic Absorption Spectroscopy (AAS)

Atomic absorption spectroscopy instruments place a sample in a high temperature flame that yields atomic species and passes selected, element specific, illumination through the flame to detect what wavelengths of light the sample atoms absorb. Either acetylene or nitrous oxide fuels the analytical flame. This process again demands that solid samples be digested (dissolved in an acid or fused with a salt) and dissolved to form a solution that can be aspirated or sprayed into the instrument's flame, while protecting the sample material from contamination or adulteration.

There are some constraints that have limited the application of AAS to general unknown analyses in forensic laboratories. The primary one is that the analyst must have some knowledge of the sample's makeup and which elements are of analytical interest so that the proper lamp can be selected to illuminate the sample. Just one or a few elements can be analyzed in any given sample run and a crucial element might be missed.

The power of AAS lays in its sensitivity and ability to detect trace elements and make it possible to distinguish otherwise identical materials. This can be a critical issue in associating an individual with a crime scene or a corporation with a damaged environmental site.

4. Graphite Furnace Atomic Absorption Spectroscopy (GFAAS)

GFAAS is similar to AAS in that the instrument measures light that the sample's atoms absorb. Samples are raised to analytical temperature in a heated graphite tube or furnace. The technique is somewhat easier and less costly to operate than flame AAS and some instruments can be set up to analyze multiple elements simultaneously, but GFAAS is about 10 to 100 times less sensitive than flame AAS.

G. Spectrophotometry (SP)

Spectrophotometry can be an extension of any of the foregoing types of spectroscopy. It is a term that refers to the quantitative analysis of spectra to compare the relative absorption or emission of different wavelengths of light. Photometry is used to determine either the absolute amount or the relative amounts of two or more compounds or elements in a sample or set of samples. In either case, SP requires the careful calibration and validation of the spectroscopy system and detector.

SP is an essential element of many forensic comparisons of spectra, especially those involving mixtures of light absorbing materials such as dyes and pigments. Fabrics and carpets are an example of materials that are often colored with a mixture of dyes and the relative proportion of dyes absorbed by a fabric may vary from dye lot to dye lot. It is necessary to establish not only that the dyes in two items are similar but also that they are present in the same amounts if the two items are to be associated.

Environmental and occupational laws, regulations, and penalties are often based not only on the presence of dangerous materials but also on the amounts of those materials that are present, making SP and essential part of an analysis.

H. Microscopical Spectrophotometry (MSP)

Microscopy has been applied to a number of spectroscopy techniques including IR, UV, VIS, and fluorescence spectroscopy that have been previously discussed. MSP is often referred to as "microspectroscopy" and involves concentrating the illuminating beam onto the sample through a microscope and then using another microscope objective to collect the radiation leaving the sample. It is not a suitable sampling system for techniques like AES, AAS, or ICP-OES that have large analytical volumes such as flames or plasmas.

MSP is especially useful when sample sizes are extremely small, including samples such as a single fiber, the edge of a multilayered paint film, a single line of ink, or an individual pigment particle. The application of spectroscopy to these types of samples has greatly increased scientists' ability to discriminate samples that are difficult or impossible to visually distinguish. This technique adds cost and complexity to the analytical instrument and forces the analyst to have a basic and functional understanding of microscopy and its effects on light beams if he or she is to produce useful results.

III. CONCLUSION

Spectroscopy offers a wide range of analytical tools for the forensic scientist's use in identifying and comparing virtually any material. The number of analytical choices and wide variety of instruments also challenge the scientist to select and apply this technology in a meaningful fashion

so the analytical results assist the courts and juries in their search for facts and truth.

SEE ALSO THE FOLLOWING ARTICLES

COLOR SCIENCE • DNA TESTING IN FORENSIC SCIENCE • GAS CHROMATOGRAPHY • INFRARED SPECTROSCOPY • LIQUID CHROMATOGRAPHY • MASS SPECTROMETRY IN FORENSIC SCIENCE • RAMAN SPECTROSCOPY • TOXICOLOGY IN FORENSIC SCIENCE

BIBLIOGRAPHY

American Society of Testing and Materials (2000). "Metals Test Methods and Analytical Procedures," Annual Book of ASTM Standards, Volume 03.06, ASTM, PA.

American Society of Testing and Materials (2000). "General Methods and Instrumentation," Annual Book of ASTM Standards, Volume 14.02, ASTM, PA.

Brode, W. (1943). "Chemical Spectorscopy," Wiley, New York.

Candler, C. (1949). "Practical Spectroscopy," Hilger and Watts, Glasgow.

Horwitz, W. (ed.) (2000). "Official Methods of Analysis of the Association of Official Analytical Chemists," Association of Official Analytical Chemists (AOAC), Washington, DC.

Humecki, H. (1995). "Practical Guide to Infrared Microspectroscopy," Mercel Dekker, New York.

Saferstein, R. (ed.) (2001). "Forensic Science Handbook, Volume I," Regents/Prentice Hall, Englewood Cliffs, NJ.

Saferstein, R. (ed.) (1993). "Forensic Science Handbook, Volume III," Regents/Prentice-Hall, Englewood Cliffs, NJ.

Siegel, J., Saukko, P., and Knupfer, G. (eds.) (2000). "Encyclopedia of Forensic Sciences," Academic Press, San Diego, CA.

Sibilia, J. (1988). "A Guide to Materials Characterization and Chemical Analysis," VCH Publishers, New York.

Williams, D., and Fleming, I. (1995). "Spectroscopic Methods in Organic Chemistry," McGraw-Hill, London.

Workman, J., and Springsteen, A. (eds.) (1998). "Applied Spectroscopy: A Compact Reference for Practitioners," Academic Press, San Diego, CA.

Thermal Analysis

David Dollimore, deceased
University of Toledo

I. Introduction
II. Differential Scanning Calorimetry and Differential Thermal Analysis
III. Thermogravimetry
IV. Evolved Gas Analysis
V. Thermomechanical Methods
VI. Modulation Techniques
VII. Simultaneous Techniques
VIII. The Role of the Computer
IX. Application of Thermal Analysis Techniques

GLOSSARY

Derivative thermogravimetry (DTG) Involves plotting the first derivative of the thermogravimetry with respect to either time or temperature. The DTG curve is plotted with the rate of mass loss on the ordinate, decreasing downward, and the temperature or time on the abscissa, increasing from left to right.

Differential scanning calorimetry (DSC) Two types of DSC are found in commercial instrumentation, namely, power-compensation DSC and heat-flux DSC.

Differential thermal analysis (DTA) Technique in which the temperature difference between a substance and a reference material is measured as a function of temperature while the substance and reference material are subjected to a controlled program.

Emanation thermal analysis Technique in which the release of radioactive emanation from a system is measured as a function of temperature while the system is subjected to a controlled temperature program.

Evolved gas detection (EGD) Technique in which the evolution of gas from a substance is noted when the system is subjected to a controlled temperature program. If the gas can be analyzed, then the term EGA is applied, and here the amount of product and its identity are measured as a function of temperature.

Isobaric mass-change determinations The equilibrium mass of a substance at a constant partial pressure of the volatile product(s) measured as a function of temperature while the substance is subjected to a controlled temperature program. The isobaric mass-change curve is plotted with mass on the ordinate, decreasing downward, and temperature on the abscissa, increasing from left to right.

Thermodilatometry (TDA) Technique in which the dimension of a substance under negligible load is measured as a function of temperature while the

substance is subjected to a controlled temperature program.

Thermogravimetry (TG) The most widely used technique of thermal analysis, in which the mass of a substance is measured as a function of temperature while it is subjected to a controlled temperature program.

Thermomechanical analysis (TMA) Technique in which the deformation of a substance under a nonoscillatory load is measured as a function of temperature while the substance is subjected to a controlled temperature program.

Thermoparticulate analysis (TPA) Technique in which the evolved particulate material in the evolved gases is measured as a function of temperature.

THE NAME *thermal analysis* is applied to a variety of techniques in which the measurement of any property of a system is recorded as the system is programmed through a predetermined range of temperatures.

I. INTRODUCTION

A. Scope of Thermal Analysis

In most laboratory experiments dealing with the properties of a material or a system the properties are measured under isothermal conditions. Separate experiments are required to measure these same property at different temperatures. In thermal analysis the specified property is measured under a controlled temperature regime. The simplest temperature regime would be that of an isothermal experiment, but in most cases the temperature is raised at a predetermined rate, for example, 10°C per minute. The interpretation then involves the variation of a particular property with both temperature and time. There is, however, a decrease in labor and time which makes such studies especially interesting for industrial applications. With more complicated temperature regimes there is an ability inherent in the method to mimic industrial processes. Industries utilizing thermoanalytical methods are listed in Table 1. The plot of the physical property of the sample recorded as a function of the temperature is said to be a *thermal analysis curve*. There is still some confusion in the literature about this name, as it was initially applied to the specific technique in which the temperature of a sample was recorded against time as it was cooled down from a particular value. The use of the name in this way persists in physical chemistry textbooks where the name *thermal analysis* is used for this specific purpose. Other conditions that have to be satisfied in the practice of thermal analysis are as follows.

TABLE I Industries Utilizing Thermoanalytical Methods

Abestos industry	Polymers, plastics, and rubbers
Industrial biochemistry	
Building industry	Pharmaceuticals
Preparation of catalysts	Medicinal
Ceramics	Organic chemicals and organic metallics
Clay processing	
Explosives and pyrotechnics	Textiles
Fats, oils, and waxes	Semiconductors
Food processing	Electronics
Fuel technology	Carbon adsorbents, charcoals, and graphites
Glass industry	
Metallurgy	Expoxy laminates
Inorganic chemical industry	Composite materials
Mineral processing	Processing and quality control
Liquid crystals	

1. The physical property and the sample temperature should be measured continuously. It should be noted that the measurement of certain properties is not easily made to comply with this particular condition.
2. In practice both the property and the temperature should be recorded automatically.
3. The temperature of the sample should be altered at a predetermined rate. In many early textbooks it is stated that the sample should be cooled or heated at a uniform rate. However, the real basis of the use of these techniques is that they should be operated on a predetermined temperature regime, as this allows various parameters to be followed. For example, one may compare directly an industrial process in which the temperature is raised, then held at a particular temperature, and then raised again.

The purpose of making the measurements is to study the physical and chemical changes which occur in a system on heating. One therefore has to interpret a thermal analysis curve by relating the property measured against temperature and interpreting the changes by noting the chemical and physical events which have taken place in the system under observation. The most obvious change in a system studied in this manner is that of mass, but calorimetry experiments predate this technique and give information concerning the enthalpy changes which take place. Evolved gas analysis and detection are also the subject of many studies where systems are heated. Another group of techniques comes under the heading of *thermomechanical analysis*. These deal with dimensional changes and with properties connected with the strength of materials when subjected to temperature changes. It

should be noted that this group must by definition include the measurement of the density of samples subjected to a programmed temperature variation. A recently introduced related technique is *dynamic mechanical analysis*, used to study the viscoelastic response of a sample under an oscillatory load. Other techniques are also used where the sample is under an oscillatory load.

A recent sophisticated approach is to modulate the temperature about a predetermined overall heating rate. Those who use these techniques must be subjected to a discipline which needs definitions and conventions so that the reader may understand the technique being utilized. In commercial equipment these techniques are often combined so that, in these simultaneous techniques, one material is subjected to two or even three measurement probes. Generally, thermal analysis techniques may be classified into three groups depending upon the way in which the physical property needs to be recorded.

1. The absolute value of the property itself can be measured, for example, the sample mass.
2. The differential method measures the difference between some property of the sample and that of a standard material, for example, their temperature difference.
3. The rate at which the property changes with temperature can be measured. These form the basis of derivative measurements and very often may be interpreted on a kinetic basis.

There exist national and international organizations that recommend nomenclature abbreviations and standards, and these organizations have set up committees that have formulated, in particular, a system of nomenclature which is adhered to in this chapter but is not always adhered to in journals and in certain other fields of science. These recommended nomenclatures and abbreviations are constantly under review, so the reader is advised to seek the most up-to-date recommendations in the literature.

B. Nomenclature

The recommended nomenclature has been put forward by the International Nomenclature Committee of the International Confederation for Thermal Analysis and Calorimetry (ICTAC). These recommendations are widely circulated in publications of the confederation and the International Union of Pure and Applied Chemistry (IUPAC). The most widely publicized report is the booklet "For Better Thermal Analysis," which is continually updated as required.

The recommended abbreviations of various techniques in thermal analysis are listed in Table II. The property measured against temperature is indicated in Table III. This list is not complete, as new techniques are continually appearing.

TABLE II Recommended Terminology for Some Thermal Analysis Techniques

Property measured	Technique name	Abbreviation
Mass	Thermogravimetry	TG
	Isobaric mass-change determination	
	Isothermal mass-change determination	
	Evolved gas detection	EGD
	Evolved gas analysis	EGA
	Derivative thermogravimetry	DTG
	Emanation thermal analysis	
	Thermoparticulate analysis	TPA
Temperature	Cooling curve[a]	
	Heating curve	
	Differential thermal analysis	DTA
Enthalpy	Differential scanning calorimetry	DSC
Dimensions	Thermodilatometry	TDA
Mechanical	Thermomechanical analysis	TMA
	Dynamic mechanical analysis	DMA

[a] Cooling curve was initially called thermal analysis. It is still called thermal analysis in many physical chemistry textbooks. Other techniques of thermal analysis measure acoustic, optical, electrical, and magnetic characteristics versus temperature. The list is not complete, as new techniques are constantly being added.

The definitions follow from Tables II and III.

Thermogravimetry (TG) is the most widely used technique of thermal analysis, in which the mass of a substance is measured as a function of the temperature while it is subjected to a controlled temperature program. The record on the thermogravimetric, or TG, curve is the mass plotted against the temperature (T) or time (t) if the variation of temperature with time can be indicated as well on the same graph. In solid-state decomposition reactions, the reactant material degrades, often to be replaced by the solid product. An example of this is the decomposition of limestone to quicklime.

In the record of the mass of the solid residue against the temperature, the decomposition of the material can be followed. This may be plotted in alternative ways, as the percentage mass loss or the fractional mass loss versus the temperature or as the fractional decomposition versus the temperature.

Derivative thermogravimetry (DTG) is not really a separate technique but involves plotting the first derivative of the TG with respect to either time or temperature. The DTG curve is plotted with the rate of mass loss on

TABLE III Property Measured versus Temperature in Some Thermal Analysis Techniques

Technique	Property measured vs temperature	Instrument
Thermogravimetry (TG) and derivative thermogravimetry (DTG)	Mass and derivative of mass with respect to temperature	Thermobalance
Differential thermal analysis (DTA)	Difference in temperature between reference and sample cell	DTA apparatus
Differential scanning calorimetry (DSC)	Enthalpy (heat) flow	DSC apparatus
Evolved gas detection and analysis (EGD and EGA)	Various properties of gas	Method varies but should always be indicated
Thermodilatometry (TDA)	Length of volume	Dilatometer
Thermomechanical analysis (TMA)	Expanion under applied stress (load)	Adapted dilatometer
Dynamic mechanical analysis (DMA)	Frequency response under oscillatory stress	DMA apparatus

the ordinate, plotted downward, and the temperature or time on the abscissa, increasing from left to right. This can be achieved by an analysis of the TG curve as a separate operation in the dedicated computer part of the equipment.

Isobaric mass-change determinations refer to the equilibrium mass of a substance at a constant partial pressure of the volatile product(s) measured as a function of temperature while the substance is subjected to a controlled temperature program.

Evolved gas analysis (EGA) is a technique in which the gas evolved from a substance subjected to a controlled temperature program is analyzed. The method of analysis should always be noted.

Emanation thermal analysis is a technique in which the release of radioactive emanation from a system is measured as a function of temperature while the system is subjected to a controlled temperature program.

Thermoparticulate analysis (TPA) is a technique in which the evolved particulate material in the evolved gases is measured as a function of temperature.

Another group of techniques involves the measurement of enthalpy changes. It has already been mentioned that, historically and in physical chemistry textbooks, the term *thermal analysis* applies to the determination of cooling or heating curves: these are techniques in which the temperature of a substance is measured as the substance is either cooled down or allowed to heat up. The cooling curve is the normal technique, and again, it may be reported as the first derivative of the heating curve with respect to time from the raw experimental data while the substance is subjected to a controlled temperature regime against time. The function dT/dt should be plotted on the ordinate, and T or t on the abscissa, increasing from left to right. It is important to state whether the technique involves a cooling process or a heating process.

The two techniques *differential thermal analysis* (DTA) and *differential scanning calorimetry* (DSC) should be considered together. DTA is a technique in which the temperature difference between a substance and a reference material is measured as a function of temperature while the substance and reference material are subjected to a controlled program. The plot is called a DTA curve; the temperature difference ΔT should be plotted on the ordinate, with the endotherm processes shown downward and the exotherm processes in the opposite direction, and the temperature or time on the abscissa, increasing from left to right. This technique is often applied quantitatively when the area of the peaks can be made proportional to the quantity of the material decomposing or to the enthalpy of the process. In this respect the equipment then serves as a calorimeter. The term DSC is applied to such experiments. There are two types of equipment in which the background temperature of the calorimeter is raised through a programmed temperature regime being imposed on the system. These are power-compensation DSC and heat-flux DSC. The method identified as power-compensation DSC was originally a copyright term employed by one of the instrument manufacturers, and often the term DSC is found in the literature applying just to power compensation equipment.

The members of the Nomenclature Committee have considered the distinction between quantitative DTA and heat-flux DSC: in their opinion, a system with a multiple sensor (e.g., a Calvert-type arrangement) or with a controlled heat leak (a Boersma-type arrangement) should be called heat-flux DSC. In practice, if the instrument manufacturers can show that the system operates as a calorimeter over a programmed temperature range, then they describe the equipment as DSC.

The usual method of plotting DTA results is, as already noted, with endothermic peaks shown downward on the plot and exothermic peaks shown in an upward direction. However, because a DSC is considered to measure thermodynamic quantities directly, the DSC plots are often found with the endothermic plots in an upward direction and the exothermic plots in a downward direction. This conforms with the IUPAC requirements for the presentation

of thermodynamic parameters. In reporting data, therefore, the directions of the endothermic peaks and the enothermic peaks should be clearly shown.

Thermodilatometry (TDA) is a technique in which the dimension of a substance under a negligible load is measured as a function of temperature while the substance is subjected to a controlled temperature program.

The record in the thermodilatometric curve is then the dimension plotted on the ordinate, increasing upward, with temperature T or time t on the abscissa, increasing from left to right. Mention should also be made of such related techniques as *linear TDA* and *volume TDA*, which differ on the basis of the dimensions measured as indicated in their names.

Thermomechanical analysis (TMA) is a technique in which the deformation of a substance under a nonoscillatory load is measured as a function of temperature while the substance is subjected to a controlled temperature program. It is used extensively in polymer studies. The mode, as determined by the type of stress applied (compression, tension, flexture, or torsion), should always be noted. As already stated, *dynamic mechanical analysis* is a technique in which the viscoleastic response of a sample under an oscillatory stress is studied while the substance is subjected to a temperature regime. *Torsional braid analysis* is a particular case of dynamic thermomechanometry where the material is supported.

Other techniques do not always find commercial instrumentation support and are usually constructed for particular, often limited, applications. An example is *thermosonimetry*, in which the sound emitted by a substance is measured as a function of the temperature while the substance is subjected to a controlled temperature program. In an associated technique, *thermoacoustimetry*, the characteristics of imposed acoustic waves are measured as a function of the temperature after passing through a substance while the substance is subjected to a controlled temperature program.

Thermoptometry is another thermal analysis technique in which an optical characteristic of a substance is measured as a function of temperature while the substance is subjected to a controlled temperature program.

Measurements of total light, light of a specific wavelength(s), refractive index, and luminescence lead to *thermophotometry*, *thermospectrometry*, *thermorefractometry*, and *thermoluminescence*, respectively. Observations using a microscope are called *thermomicroscopy*. The microscopy observations, however, are often referred to under the general term of hot-wire microscopy. *Thermoelectrometry* is a technique in which an electrical characteristic of the substance is measured as a function of temperature. The most common measurement here is resistance, conductance, or capacitance.

Thermomagnetrometry is yet another technique of thermal analysis in which the magnetic susceptibility of a substance is measured as a function of temperature while the substance is subjected to a controlled temperature program.

Sometimes more than one technique is used in an investigation. Such multiple techniques can be classified as follows. The term *simultaneous techniques* covers the application of two or more techniques to the same sample at the same time, for example, simultaneous TG and DTA. This is becoming a widely used practice and is discussed in more detail in Section VII.

C. Symbols

The abbreviations for each technique have already been noted (see Tables II and III). In polymer studies, however, the distinction between T_g and TG may cause confusion. Here the abbreviation TG refers to thermogravimetry, while T_g represents the glass transition temperature. This has caused a number of investigators and instrument manufacturers to use TGA for TG to avoid confusion. Other aspects of the use of symbols are mentioned in the following list.

1. The international system of units (SI) should be used wherever possible.
2. The use of symbols with superscripts should be avoided.
3. The use of double subscripts should also be avoided.
4. The symbol T should be used for temperature whether expressed as degrees Celsius (°C) or as kelvins (K). For temperature intervals the symbol K or °C can be used.
5. The symbol t should be used for time, whether expressed as seconds (s), minutes (min), or hours (h).
6. The heating rate can be expressed either as dT/dt when a true derivative is intended or as β in K min^{-1} or °C min^{-1}. The heating rate so expressed need not be constant over the whole temperature range and can be positive or negative, so this should be stated.
7. The symbols m for mass and W for weight are recommended.
8. The symbol α is recommended for the fraction reacted or changed.
9. The following rules are recommended for subscripts:
   a. Where the subscript relates to an object, it should be a capital letter, e.g., M_s represents the mass of the sample, and T_R represents the temperature of the reference material.
   b. Where the subscript relates to a phenomenon occurring, it should be lowercase, e.g., T_g

represents the glass transition temperature, T_c represents the temperature of crystallization, T_m represents the melting temperature, and T_t represents the temperature of a solid-state transition.

c. Where the subscript relates to a specific point in time or a point on the curve, it should be a lowercase letter or a number, e.g., T_i represents the initial temperature, $t_{0.5}$ represents the time at which the fraction reacted is 0.5, $T_{0.3}$ represents the temperature at which the fraction reacted is 0.3, T_p represents the temperature of the peak in DTA or DSC, and T_c represents the temperature of the extrapolated onset. This can also be applied to DTG techniques.

D. Standardization

No single instrument design or set of experimental conditions is optimum for all studies. The techniques are dynamic in nature and flexible in use but produce data which may be highly dependent upon the procedure. The prime requirements for standardization can be listed as follows.

1. The provision of a common basis for relating independently acquired data.
2. The provision of the means for comparing and calibrating all available instrumentation, regardless of design.
3. The provision of the means for relating thermoanalytical data to physical and chemical properties determined by conventional isothermal procedures.

This means in effect that the geometry of the measuring system and its effects on experiments must be noted. One of the most obvious points regarding standardization is the use of proper materials to establish the validity of the thermocouple readings. Such standards would then reflect the effect of the experimental design (i.e., the geometry of the instrumentation layout), and in this way a correction can be applied to the recorded temperature. In some units, for example, the temperature measuring device is used to control the furnace temperature and is located away from the sample. This means that the sample or system under observation has a temperature record in the thermal analysis curve which is not actually the temperature of the sample; again, this must be corrected. By including curves of standard materials obtained under particular conditions, it is possible to relate and estimate such errors. It is necessary to impress on instrument manufacturers the need to locate the temperature measuring and control device as close to the sample as conveniently possible.

TABLE IV Materials that Can Be Used for Temperature Calibration in Thermal Analysis[a]

Material	Transition type	Peak temperature (°C)
Polystyrene	Glass transition	~101
1-2 Dichloroethane	m.p.	−32
Cyclohexane	Transition point	−83
	m.p.	+7
Phenyl ether	m.p.	30
o-Terphenyl	m.p.	58
Potassium nitrate	Transition point	128
Indium	m.p	157
Tin	m.p.	232
Potassium perchlorate	Transition point	300
Silver sulfate	Transition point	430
Quartz	Transition point	573
Potassium sulfate	Transition point	583
Potassium chromate	Transition point	665
Barium carbonate	Transition point	810
Strontium carbonate	Transition point	925

[a] These temperature calibration materials are supplied by instrument manufacturers on request. The essential condition is that the material should be pure. The above data are generally available for determination of temperature on DTA or DSC equipment.

Most modern equipment has the temperature measuring device located very close to the samples. The instrument manufacturers will supply suitable reference materials on request with appropriate certification. Table IV sets out *some* suitable materials that can and have been used as temperature standards for calibration purposes. Directly these are most useful for calibrating DTA or DSC units. In DSC units knowledge of the enthalphy changes may also be required. Table V sets out the enthalpy of fusion for selected materials. Again, most DSC instrument manufacturers provide materials with their equipment for this purpose. Indium is used as a calibration material in many DSC units, and other systems undergoing phase changes can be used in a similar way.

There is a need to establish the "proper" temperature in TG experiments. Here the use of materials with magnetic transitions that can be displayed on a mass-loss curve and be referred to the temperature prove to be most useful. The reference sample is placed in the sample container and suspended within a magnetic field gradient. The magnet applying the field can be either a permanent magnet or an electromagnet which can be placed, for the purpose of calibration, near the same location. At the reference material's Curie temperature the magnetic effect diminishes to zero and the TG unit indicates an apparent mass change.

TABLE V Enthalpies of Fusion for Selected Materials

Material	Melting point (°C)	Enthalpy of fusion (J g^{-1})
Naphthalene	80.3	149
Benzoic acid	122.4	147.4
Phenacetin	134.7	173.0
Indium	156.6	28.5
Tin	231.9	59.5
Bismuth	271.3	53.1
Lead	327.5	23.1
Zinc	419.6	112.0
Aluminum	660.4	399.4
Sodium chloride	800.0	495.0
Silver	961.9	107.0
Gold	1064.4	64.0

Table VI indicates suitable materials recommended for use as magnetic standards. McGhie and co-workers have introduced another method for calibration of TG balance termed *fusible link temperature calibration*. In this method loops of metal are attached to the sample part of the TG balance. At their melting point these links condense to a liquid and fall off the balance (to be collected in a suitable container). The melting point is thus recorded as a sudden loss in weight. Table VII lists suitable materials for this method.

E. Reporting Thermal Analysis Data

The committee on standardization has also reported the manner in which information obtained on thermal analysis equipment should be published. Its recommendations are as follows.

To accompany each DTA, TG, EGA, and EGD or thermochemical record, the following information should be provided.

1. Substances should be identified (sample, reference, diluent) by a definite name and empirical formula or with equivalent composition data.

TABLE VI Magnetic Transition Temperatures Using ICTAC-Certified Magnetic Reference Materials GM 761

Material	Transition temperature (°C)
Permanorm 3	259.6 ± 3.7
Nickel	361.2 ± 1.3
Mumtal	403.0 ± 2.5
Permanorm 5	431.3 ± 1.6
Trafoperm	756.2 ± 1.9

TABLE VII Fusible Link Method for Calibrating a TG Unit

Material	Temperature (°C)		
	Observed	Corrected	Literature
Indium	159.90 ± 0.97	154.20	156.63
Lead	333.02 ± 0.91	331.05	327.50
Zinc	418.78 ± 1.080	419.68	419.58
Aluminum	652.23 ± 1.32	659.09	660.37
Silver	945.90 ± 0.52	960.25	961.93
Gold	1048.70 ± 0.87	1065.67	1064.43

2. A statement of the sources of all substances should be given, with details of their history, pretreatment, and chemical purity as far as these are known.

3. Measurement of the average rate of linear temperature change over the temperature range involved should be reported. Nonlinear temperature programs should be described in detail.

4. Identification of the sample atmosphere by pressure, composition, and purity is important and should be recorded: in particular, it should be stated whether the atmosphere was static, self-generated, or dynamic and whether it passed through or over the sample. Where applicable the ambient atmospheric pressure and humidity should also be specified. If the pressure is other than atmospheric, full details of the method of control should be given.

5. A statement of the dimension, geometry, and materials of the sample holder should be provided.

6. A statement of the method of loading should also be provided where this is applicable.

7. The abscissa scale should be identified in terms of time or temperature, and the location at which the temperature was measured should be provided. Time or temperature should be plotted so that it increases from left to right.

8. A statement of the method used to identify intermediates or final products should be given.

9. A faithful reproduction of all the original records should be available. With some instruments this is quite difficult, for the record is of mass versus time and temperature versus time, provided as two separate plots. In most modern units, however, this kind of problem can be avoided because the data are logged into a dedicated computer and automatically printed out as a graph, and the computer program then takes care of this kind of problem, giving the direct relationship between the measurement of the property selected and the temperature.

10. Identification of the apparatus used by type and/or commercial name is essential, together with details of the location of the temperature measuring thermocouples as already indicated. Again, this is important, as people using

equipment very often modify the commercial units to suit their particular needs.

There are, of course, other data that should be reported for specific techniques, and these are mentioned when dealing with these techniques.

F. Components of Thermal Analysis Instrumentation

The thermal analysis equipment enables the sample to be heated at a predetermined rate so that its temperature and one or more of its physical properties can be continuously measured and recorded. There are, therefore, three basic units involved: first, the measuring unit; second, the temperature control unit; and third, the recording unit. The simple arrangement of these components is given in Fig. 1. There is a measuring unit in all these systems and this must be in a particular position in the furnace. There must also be a system for passing a controlled atmosphere around the sample and a thermocouple proximate to the sample so that the temperature is accurately measured. The sample atmosphere is very important in liquid gas systems or in chemical reactions, and therefore, the measuring unit should be capable of operating under inert reactive gas or vacuum conditions. One point that is missed by most instrument manufacturers is the fact that the so-called vacuum systems require continuous pumping to eliminate the leak of gas into the system through various parts of the equipment. Units which hold their vacuum or which operate at controlled pressures other than atmospheric are built only by specialist instrumentation companies.

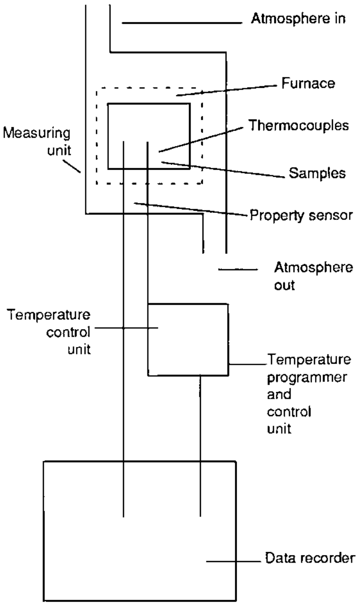


FIGURE 1 Component parts of a thermal analysis unit.

The temperature control unit can be a simple furnace and programmer. The fact that in some cases the instrumentation operates a control of the furnace temperature, rather than the sample temperature, is an important design feature which has already been noted. The recording unit receives a signal from the measuring unit and the temperature thermocouple. In most cases this is fed into a dedicated computer, and this allows the signals to be displayed in a variety of different ways, depending on the requirements of the operator.

In general, thermal analysis instruments may be divided into two groups: differential instruments, which contain the sample and a reference cell in similar environments and provide a difference signal of the properties; and derivative instruments, which note the change in the property signal versus the temperature as a proper derivative signal. However, it must be noted that the latter type of device is no longer required, as the property signal can be fed into a dedicated computer and the dedicated computer can now do the derivative calculations and provide a curve of both the property versus the temperature and the rate of change of that property versus the temperature.

1. Furnace Design

Furnaces have to be designed for particular applications since they have to be compatible with the measuring system and also with problems associated with each technique, for example, convection currents in TG. However, there are some general observations that can be made, first, regarding the thermal capacity of the furnace. If a large furnace is used, then one is going to have a range within the furnace at which a uniform temperature can be recorded. Small furnaces, however, will not have this uniform range of temperature, and the positioning of the sample in smaller furnaces becomes quite important. The second point about the size of the furnace is that large furnaces will take considerable time to reach a particular temperature and also take some time to cool down. The smaller the furnace, the easier it is to cool the temperature of the furnace back to ambient. The sample and reference material in differential measurements must also be subjected to the same temperature change. This generally involves a design feature involving both the unit in which the measurements are made and the furnace itself. One further point about furnace size and shape is that a long narrow furnace will generally give a larger uniform hot zone than a short wide furnace of similar volume. All instrumentation furnaces are electrically powered (although high-frequency inductive and infrared heating furnaces have been put on the market). The resistance wire is generally coiled around an insulating packing. The outside of the furnace is generally well insulated, although this has a bearing on the rate at

TABLE VIII Upper Temperature Limits for Common Furnace Resistance Elements

Furnace winding(s)	Temperature limit (°C)	Atmosphere[a]
Nichrome	1000	A
Chromel A	1100	A
Tantalum	1330	B
Kanthal	1350	A
Platinum	1400	C
Platinum–10% rhodium	1500	C
Platinum–20% rhodium	1500	C
Kanthal super	1700	A
Rhodium	1800	C
Molybdenum	2200	D
Tungsten	2800	D

[a] A—an oxidizing atmosphere can be used (oxygen or air); B—a nonoxidizing atmosphere can be used (inert or vacuum); C—in these cases oxygen or air can be used at lower temperatures, but at higher temperatures an inert atmosphere is recommended; D—hydrogen should be used.

which the furnace will cool down. Because of the danger of a magnetic field interfering with the measurements of certain physical properties of the sample or system, the furnace should be noninductively wound: that is, it should have two similar windings carrying current in opposite directions so that their magnetic fields cancel, and also, the spacing of the windings should be decreased toward the end of the furnace to compensate heat losses. These points are made because, for special purposes, the operator might well require the construction of furnaces to a specific design. The resistance elements of the furnace control the temperature that can be obtained. Table VIII gives a range of resistance elements which, in theory, allows a temperature of 2800°C to be reached. Furnaces operating above a temperature of 1350°C should do so in an inert or reducing atmosphere. Graphite or silicon carbide resistance bars are used instead of metal bars in furnaces in which temperatures can go over 2000°C, and these are simple to operate, provided the restrictions on atmosphere control are noted.

The furnace should generally be mounted so that it can be moved easily and also put back in exactly the same position each time to minimize any errors due to the geometry of the system being altered by the movement which is necessary to load the sample in and to generally inspect the equipment. Where large furnaces are used, it is essential to have some kind of cooling system to allow the furnace to be programmed rapidly back to ambient temperature.

2. Temperature Programmers

Most thermal analysis results require a simple linear heating program, so that a typical instrument will allow programming from ambient to a preset maximum temperature followed by the equipment being either switched off or programmed to cool. In some units natural cooling is used; in others there is the additional choice of maintaining the maximum temperature isothermally. The programmed linear cooling process can often be allowed to go to a preset minimum temperature, followed by the equipment being switched off, by programmed heating, or by maintenance this minimum temperature isothermally. In most modern equipment, cycling at a given rate between two preset temperatures can be achieved. It has also been found to be convenient to operate the equipment isothermally at a rapidly preset temperature or to temperature jump so that a number of isothermal intervals for a predetermined period are imposed upon the system or the sample. In other units very rapid quench cooling from a preset temperature can be achieved. It should be noted that the isothermal treatment of materials does not fall within the obvious definition of thermal analysis unless this is taken to mean that the material is heated at a zero rate, but most operators find it important because they wish to refer their heat treatment at a determined rate against more classical studies involving isothermal operation. In other applications of the work, for example, simulation of industrial processes, more complicated program temperatures are required, and with the aid of computer programming, this is generally possible. Since most thermocouples need a reference cold junction, this is now generally provided in the form of an electronic ice point reference unit.

3. Recording Systems

In the current design of thermal analysis equipment the recording system is such that the signals are received via the computer system. These signals can be obtained in the form a of a digital readout or in the form of an X–Y plot. These data can be provided using software provided by the instrument manufacturer, by the dedicated computer manufacturer, or by the laboratory using the equipment. This enables the material to be subjected on the spot to various analytical processes which indicate at once the observations and the changes taking place.

G. Publications and Books on Thermal Analysis

The literature on thermal analytical methods is widely scattered because it is generally abstracted under the application heading rather than under the method heading. The American Chemical Society publishes its Chemical Abstracts, which embrace most published research papers, and all the articles which contain thermal analysis data then appear in *C.A. Select Abstracts on Thermal Analysis*. There are two international journals that publish papers on

thermal analysis, *Journal of Thermal Analysis* and *Thermochimica Acta*. Other journals, however, have appeared which contain an increasing number of papers on thermal analysis; two of them are *Journal of Materials Research* and the *Journal of Analytical and Applied Pyrolysis*. The Japanese *Journal of Calorimetry and Thermal Analysis* contains material which is pertinent, and the various newsletters issued by national thermal analysis societies also contain material of interest (e.g., *ICTA Newsletter*, *Bulletin de l'Association Francaise de Calorimetrie et d'Analyse Thermique*, *NATAS Notes*, *Aicat Notizie*). These socieity publications serve a useful purpose by carrying information on recent developments, meetings, books, and so on that may be of interest. Every few years the proceedings of the International Conference on Thermal Analysis and Calorimetry provides an up-to-date picture of the stages of development of thermal analysis in different fields of science. These proceedings have been appearing since 1965. The journal *Analytical Chemistry* provides a 2-yearly review of highlights in the field of thermal analysis.

II. DIFFERENTIAL SCANNING CALORIMETRY AND DIFFERENTIAL THERMAL ANALYSIS

A. Introduction

In a group of techniques the changes in the heat content (enthalpy) or the specific heat of a sample are noted with respect to the temperature. The terms have been defined previously, but it is best to recall that differential thermal analysis (DTA) is a technique in which the temperature difference between a substance and a reference material is measured as a function of temperature while the substance and the reference material are subjected to a controlled temperature program. The "classical" DTA instrument is shown in Fig. 2. The record is the DTA curve, with the temperature difference (ΔT) plotted on the ordinate, with en-

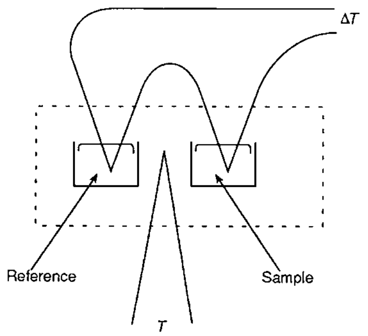


FIGURE 2 Schematic diagram of a DTA unit.

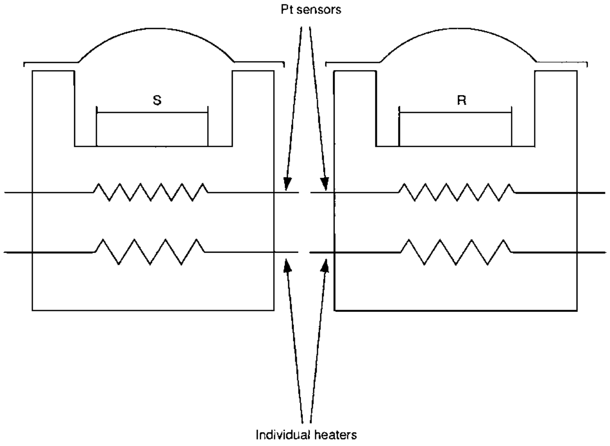


FIGURE 3 Schematic diagram of a power-compensated DSC unit.

dothermic reactions downward, and T or t on the abscissa, increasing from left ot right. The technique can be made quantitative by calibration and with the correct design of cells. Such units are labeled by instrument manufacturers as differential scanning calorimetry (DSC). The name, however, was reserved for a long time for a special method of obtaining calorimetric results, namely, by neutralizing the ΔT signal with auxiliary heaters and recording the energy required to neutralize the signal. Such equipment is shown schematically in Fig. 3 and would now be called a "power-compensated DSC." The quantity measured is the rate of change of enthalpy plotted against temperature. As already noted, the IUPAC thermodynamic convention is then to plot exothermic quantities (where the system loses energy) in a downward direction and endothermic quantities (where the system gains energy) in an upward direction. Although this is opposite to the recommended plotting of the ΔT signal, in DTA there is no real conflict, for ΔT is logically a drop in temperature for an endothermic process and a gain in temperature for an exothermic process. It is, of course, necessary to indicate in which direction endothermic and exothermic processes are shown on DTA or DSC plots. The alternative to power compensation DSC is heat-flux DSC, where the cells are designed so that the heat flux can be accurately used in the calibration process. Such a cell is shown schematically in Fig. 4. In a calibrated unit the enthalpy change in any process is proportional to the area under the curve.

B. Nomenclature for DTA

It should be noted that in DTA two cells are used, one containing the sample under investigation and the other a reference material of inert behavior. The two cells are

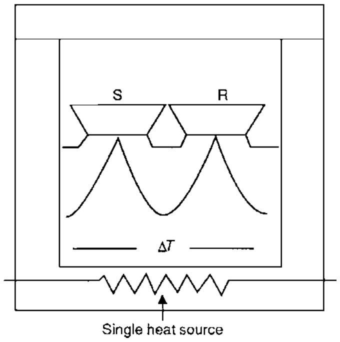


FIGURE 4 Schematic diagram of a Boersma-type DSC unit.

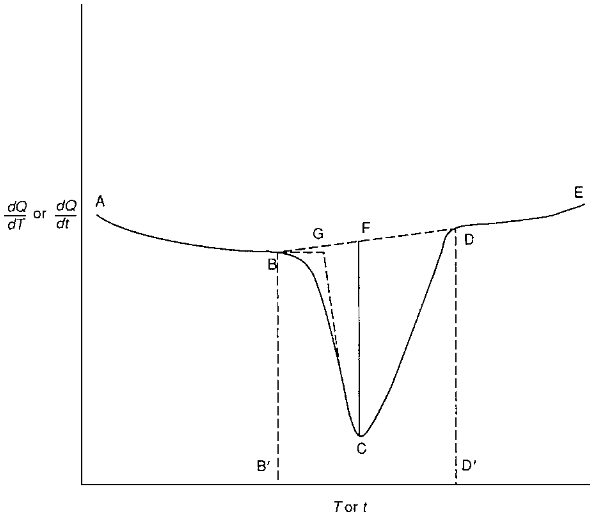


FIGURE 5 Formalized DSC signal. Note that in DTA the signal is ΔT instead of dQ/dT.

heated in a furnace subjected to a programmed increase or decrease in temperature. Two thermocouples, one in each cell, are "backed off" one against the other, and the difference in temperature (ΔT) between the two cells is plotted against time or temperature.

Nomenclature advice specific to DTA is as follows. The *sample* is the actual material investigated, whether diluted or undiluted. The *reference material* is a known substance, usually inactive thermally over the temperature range of interest. The term *inert material* is often used but is not recommended by the ICTA. The *specimens* are the sample and reference material. The *sample holder* is the container or support for the sample. The *reference holder* is the container or support for the reference material.

The *specimen-holder assembly* is the complete assembly in which the specimens are housed. Where the heating or cooling source is incorporated in one unit with the containers or supports for the sample and reference material, this would be regarded as part of the specimen-holder assembly. A *block* is a type of specimen-holder assembly in which a relatively large mass of material is in intimate contact with the specimens or specimen holders.

The *differential thermocouple*, or ΔT thermocouple, is the thermocouple system used to measure the temperature difference. In both DTA and TG the *temperature thermocouple* is the system used to measure temperature; its position with respect to the sample should be stated.

The *heating rate* is the rate of temperature increase (degrees per minute); likewise, the cooling rate is the rate of temperature decrease.

In DTA it must be remembered that although the ordinate is conventionally labeled ΔT, the output from the ΔT thermocouple will in most instances vary with temperature, and the measurement initially recorded is the e.m.f. output, E—that is, the conversion factor, b, in the equation $\Delta T = bE$ is not constant since $b = F(T)$, and a similar situation occurs with other sensor systems. The data fed into the computer work station, however, are corrected so that the plot presented is that of ΔT versus T.

All definitions refer to a single peak such as that shown in Fig. 5. Multiple-peak systems, showing shoulders or more than one maximum or minimum, can be considered to result from the superposition of single peaks.

The baseline (AB and DE) corresponds to the portion or portions of the DTA curve for which ΔT is changing only slightly.

A peak (BCD) is that portion of the DTA curve which departs from and subsequently returns to the baseline.

An endothermic peak or endotherm is a peak where the temperature of the sample falls below that of the reference material; that is, ΔT is negative.

An exothermic peak, or exotherm, is a peak where the temperature of the sample rises above that of the reference material; that is, ΔT is positive.

The peak width (B′D′) is the time or temperature interval between the point of departure from and the point of return to the baseline. There are several ways of interpolating the baseline, and that given in Fig. 5 is only an example. The location of points B and D depends on the method of interpolation of the baseline.

The peak height is the distance, vertical to the time or temperature axis, between the interpolated baseline and the peak tip (C, in Fig. 5).

The peak area (BCDB) is the area enclosed between the peak and the interpolated baseline.

The general code of practice for recommending thermal analysis should be supplemented as follows for DTA.

1. Wherever possible, each thermal effect should be identified and supplementary supporting evidence shown.
2. The sample weight and dilution of the sample should be stated.

3. The geometry and materials of thermocouples and the location of the various thermocouples should be given.
4. The ordinate scale should indicate the deflection per degree Celsius at a specified temperature.

C. Position of the "Temperature" Thermocouple

It is normal to program the temperature to produce a linear heating rate. This, however, is not a sufficiently precise statement. One can use either the thermocouple in the reference material or the thermocouple in the sample as the indicator of temperature. In many systems the sample thermocouple is advocated as the indicator of the temperature. In the model set up by Cunningham and Wilburn there is a separate thermocouple to generate the T signal and two further thermocouples to give the ΔT signal. Using this model the same authors came to the conclusion that the temperature-measuring the thermocouple junction is to be placed in the sample. This point is also made by Mackenzie an Mitchell (1970) with reference to the above theoretical work of Wilburn and that of Grimshaw *et al.* (1945).

Two points can now be made. First, although it is generally recommended that the sample cell is the proper location for the temperature thermocouple, this means that a linear response is possible only while no reaction or phase change takes place. This is demonstrated in Fig. 6. Second, the peak height of the ΔT signal varies with the temperature at which T is measured.

In practice it is often found that the factors relating to the sample which are important are the particle size, the particle size distribution, the packing of the sample, the dilution of the sample with reference material (although most modern units dispense with this), and control of the atmosphere. The dilution of the sample represents an attempt to make the thermal characteristics of the two cells identical. There is, however, the possibility of reaction between the diluent reference material and the sample under investigation. However, as noted above, in most modern units dilution is not necessary and the reference cell is used empty. This is because the high sensitivity of the equipment requires only a few milligrams of sample to be studied.

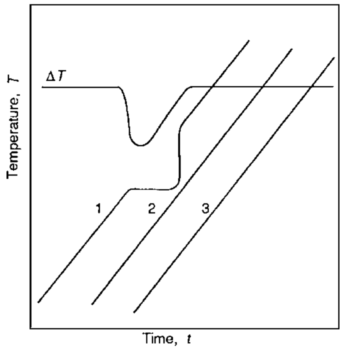


FIGURE 6 DTA plots: (1) T, sample cell; (2) T, reference cell; (3) T, third thermocouple. Scale displaced in 2 and 3.

Most commercial firms issue concise operating instructions and many application briefs. These should be referred to as often as needed before a deeper search of the literature is attempted. The possibility of using a single-cell model in the future is real, for the temperature of the "inert" could be calculated by some form of computer program and subtracted from the single-cell temperature reading to given the ΔT signal.

D. Theory of DTA

The main applications of DTA are (i) to describe the thermal decomposition and transitions occurring on heating a material through a programmed temperature range, (ii) to measure the heat of reaction, and (iii) to determine the kinetic parameters. The first is reasonably clear, the second requires an explanation, and the third, once the energy terms have been related to the weight changes, is similar to the treatment of kinetic data on the thermobalance.

In dealing with the measurement of the heat of reaction, the theories of DTA can be placed in two categories: (i) those which deal with heat transfer alone and (ii) those which deal with the reaction equation, that is, take into account the chemical nature of the reaction.

A simplified heat transfer theory is based on the method developed by Vold (1949).

The equation of heat balance for the cell containing the reaction is

$$C\,dT_1 = dH - K(T_3 - T_1)\,dt, \qquad (1)$$

and that for the cell containing the inert material is

$$C\,dT_2 = -K(T_3 - T_2)\,dt, \qquad (2)$$

where C is the heat capacity of each cell; T_1, T_2, and T_3 are the temperatures of the reactant, reference material, and block, respectively; K is the heat transfer coefficient between the block and the cell; and dH is the heat evolved by the reaction in time dt.

In writing these equations two assumptions are made: (i) the heat capacities of the two cells are the same and do not change during the reaction, and (ii) there is a uniform temperature throughout the sample at any instant. Otherwise a single value (T_1 and T_2) could not be written for the temperature of the material in the two cells.

Subtracting Eq. (2) from Eq. (1) gives

$$C(dT_1 - dT_2) = dH + K(T_1 - T_2)\,dt.$$

The ΔT signal is, in fact, $T_1 - T_2$, so

$$Cd\Delta T = dH - K\ \Delta T\ dt. \tag{3}$$

To determine the total heat of reaction it is necessary to integrate from $t = 0$ to $t = x$:

$$\Delta H = C(\Delta T - \Delta T_0) + K\int_0^x \Delta T\ dt,$$

when $t = 0$, $T = 0$, and when $t = x$, $T = 0$:

$$\Delta H = K\int_0^x \Delta T\ dt \tag{4}$$

or

$$\Delta H = KS,$$

where S is the area of the DTA peak.

The influence of physical properties on the baseline can be considered more realistically by assigning different values of C and K to each cell and considering the simple case where there is no reaction, that is, $dH = 0$; then

$$C_1\, dT_1 = K_1(T_3 - T_1)\, dt$$

and

$$C_2\, dT_2 = K_2(T_3 - T_2)\, dt.$$

C_1 and K_1 refer to the reactant cell, and C_2 and K_2 to the reference cell. Rearrangement gives

$$T_1 = T_3 - \frac{C_1\, dT_1}{K_1\, dt}$$

and

$$T_2 = T_3 - \frac{C_2\, dT_2}{K_2\, dt}.$$

dT_1/dt and dT_2/dt represent heating rates and should be identical, that is.

$$\frac{dT_1}{dt} = \frac{dT_2}{dt} = \frac{dT}{dt}.$$

Then

$$\Delta T = T_1 - T_2 = \frac{dT\,(K_1C_2 - K_2C_1)}{dt\,(K_2K_1)} \tag{5}$$

We now have three cases.

$$\frac{C_2}{K_2} = \frac{C_1}{K_2} \quad \text{and} \quad Ck \neq f(T).$$

this is demonstrated by the zero value of ΔT (Fig. 7).

$$\frac{C_2}{K_2} > \frac{C_1}{K_1} \quad \text{and} \quad Ck \neq f(T),$$

which gives a positive constant value of ΔT (Fig. 7).

$$\frac{C_2}{K_2} < \frac{C_1}{K_1} \quad \text{and} \quad Ck \neq f(T),$$

which gives a negative constant value of ΔT (Fig. 7).

If

$$\frac{C_2}{K_2} < \frac{C_1}{K_1} \quad \text{but} \quad Ck = \text{linear } f(t)$$

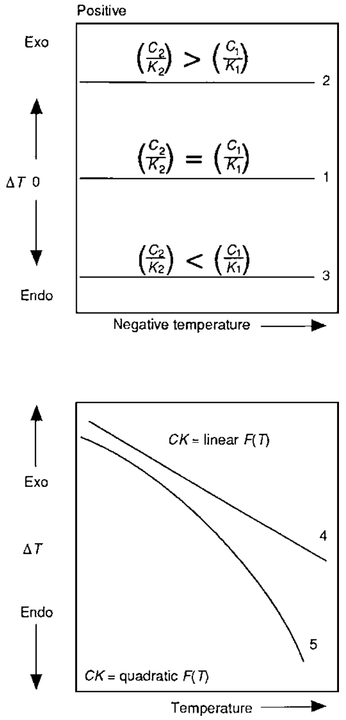


FIGURE 7 DTA baseline behavior according to Vold. Note that in the top plot CK is not a function of temperature; and in the bottom plot (C_2/K_1).

or if

$$\frac{C_2}{K_2} < \frac{C_1}{K_1} \quad \text{but} \quad Ck = \text{quadratic } f(t),$$

then the curved plot in Fig. 7 results. If, however, $C_2/K_2 > C_1/K_1$, the slopes are in the opposite direction.

For a more comprehensive treatment the publications of Wilburn and his coauthors should be consulted.

The practical tests of the use of DTA equipment in this way are to check whether the peak area is proportional to the quantity of material under examination and also to check the area under the peak for materials of a known heat of reaction. If the equipment responds properly to calibration tests of this kind, then it would seem, within the limits of the calibration range, to be proper to use it as a scanning calorimeter.

III. THERMOGRAVIMETRY

A. Introduction

As already noted, a *thermobalance* is an apparatus for weighing a sample continuously while it is being heated or cooled. The *sample* is the actual material investigated, whether diluted or undiluted. The *sample holder* is the container or support for the sample. Again, as already mentioned, in thermal analysis certain experimental items

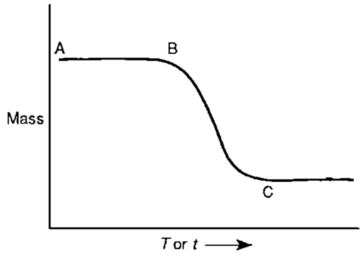


FIGURE 8 Formalized TG curve.

should be reported. The following additional details are also necessary in the reporting of TG data.

A statement must be made concerning the mass and the mass scale for the ordinate. The actual weight of the sample is reported, so a downward trend represents a mass loss. Additional scales, for example, mass loss, fractional decomposition, and molecular composition, can also be used according to the calculations which are subsequently to be made.

If derivative TG is employed, the method of obtaining the derivative should be indicated and the units of the ordinate specified.

The definitions used in TG can be illustrated by reference to the single-stage process illustrated in Fig. 8. Reference to multistage processes can be considered by thinking of the multistage process as being a series of single-stage processes.

A *plateau* (AB in Fig. 8) is that part of the TG curve where the weight is essentially constant.

The initial temperature, T_i (B in Fig. 8) is the temperature (on the Celsius or Kelvin scales) at which the cumulative weight change reaches a magnitude that the thermobalance can detect.

The final temperature, T_t (C in Fig. 8) is that temperature at which the cumulative weight change reaches a maximum. The reaction interval is that temperature difference between T_f and T_i as defined above.

B. Design Factors

The basic instrumental requirements for TG are a precision balance, a furnace capable of being programmed through a required regime of temperature change, and a computer workstation capable of programming the furnace, recording the weight change, and processing the data (see Fig. 9).

The essential requirements of an automatic and continuously recording balance are similar to those of an analytical balance and include accuracy, sensitivity, reproducibility, and capacity. In addition, a recording balance should have an adequate range of automatic weight adjustment, a high degree of mechanical and electronic stability, and a rapid response to weight changes and be unaffected by vibration.

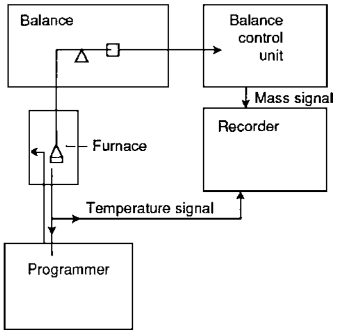


FIGURE 9 Schematic diagram of a thermobalance and control unit.

Two weighing systems need to be noted, namely, deflection and the null-point balances. There is a variety of deflection balances that can be constructed—beam type, helical spring, cantilevered beam, torsion wire, etc.—but they suffer in that the sample under observation will not remain in a fixed position in the furnace. For this reason most units employ a null-point balance.

In the use of null-point balances, a sensor must be used to detect the deviation of the balance beam from its null position, and a variety of methods may be used to detect deviations from the horizontal or vertical norm. One common system would is make use of the varying intensities of a light source impinging upon a photoelectric cell. The usual arrangement incorporates a light source, a shutter or mirror, and either single or double phototubes. The displacement of the shutter attached to the balance beam (or spring) intercepts the light beam, thus either increasing or decreasing the light intensity acting on the phototube. The resulting change in current magnitude from the phototube is then used to restore the balance to its null point.

Furnace design features and temperature measurement have already been discussed.

IV. EVOLVED GAS ANALYSIS

A. Introduction

In EGA the gases evolved from the decomposition of materials in a thermal analysis unit are analyzed. There are many methods of analyzing gases. In the past, specific chemical methods have been favored, but now instrumental methods based largely on mass spectrometric methods, chromatography, or infrared spectroscopy are generally practiced. In certain applications, however, specific chemical analysis is still used. In thermal analysis there must be an interface between the heat-treated sample and the gas detection unit. Gas analysis is rarely used in such cases by itself but more commonly combined with TG or DTA.

In an earlier section details of reporting thermal analysis were noted. In EGA, the following additional details are necessary.

1. The temperature environment of the sample during reaction should be clearly stated.
2. The ordinate scale should be identified in specific terms where possible. In general, increasing concentration of evolved gas should be plotted upward. For gas density detectors, increasing gas density should also be plotted upward.
3. The flow rate, total volume, construction, and temperature of the system between the sample and the detector should be given, together with an estimate of the time delay within the system.
4. Location of the interface between the systems for heating the sample and detecting or measuring evolved gases should be noted.
5. In the case of EGA, when the extact units are not used, the relationship between the signal magnitude and the concentration of species measured should be stated. For example, the dependence of the flame ionization signal on the number of carbon atoms and their bonding as well as on the concentration should be given.

B. Methods of Analysis

The methods of gas analysis are numerous, and any attempt at classification cannot be complete. With regard to present-day use, however, the following methods may be cited.

1. Chemical analysis, usually based on absorption of the gas or a component of the gas mixture in solution.
2. Mass spectrometric methods.
3. Infrared spectroscopy.
4. Gas chromatography.

Applied to thermal analysis there is usually the problem to consider as to how best to interface the gas detector or analysis unit with the rest of the equipment. There are also problems which may arise when the data are used for some specialized purpose, for example, to obtain kinetic data. The method of chemical analysis is rarely practiced in commercial units.

The use of an appropriate interface to sample the gas stream coming from the material under heat treatment must usually conform to the layout shown schematically in Fig. 10.

The following particular points about interfacing should be noted.

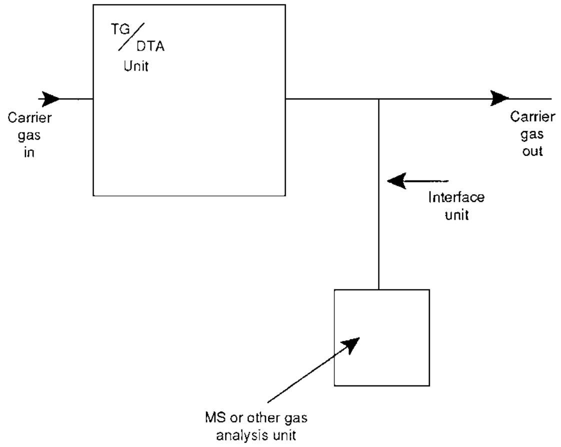


FIGURE 10 Schematic diagram showing the use of an interface in evolved gas analysis.

1. The thermal analysis unit may operate at 1 atm, but some gas analysis units, particularly mass spectrometers, operate at a very high vacuum.
2. Diffusion in the interface device may be a complicating factor.
3. Some problems of interfacing may be applicable to one particular method of gas analysis, and rather than attempt to solve a very difficult problem, it may be more convenient to choose another form of gas detector.
4. To eliminate side reactions it is necessary to put the interface as close as possible to the decomposing sample.
5. Carbon monoxide presents problems in mass spectrometry. This is because nitrogen has the same mass number as carbon monoxide, and in most mass spectrometers "ghost" peaks of nitrogen may be present. The solution is probably to use infrared detector devices for analysis of carbon monoxide.
6. Water presents problems of analysis. This is mainly because of its persistent adsorption and the difficulty in degassing it. This problem is apparent in the use of mass spectrometers and also in gas chromatography.

Prout and Tompkins (1946) devised a simple method for measuring the pressure of gas produced from a decomposing solid. They kept the total pressure very low, which allowed the kinetics of decomposition to be determined. The equipment is shown in Fig. 11, and it has been successfully applied to decomposition of potassium permanganate.

C. Mass Spectrometric Methods

In Table IX a distinction is made between *mass spectrometric thermal analysis*, in which the sample is actually located in the mass spectrometer, and *mass spectrometry coupled to either DTA, TG, or both*. The latter type is most often used by commercial instrument manufacturers. The

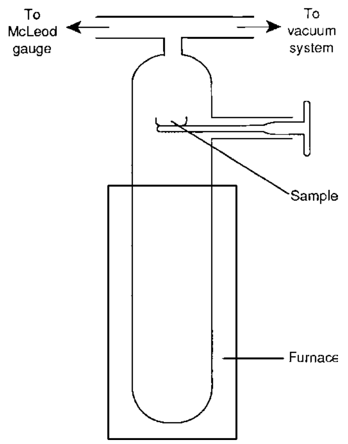


FIGURE 11 A Prout and Tompkins unit.

main problem is that of interfacing, as the mass spectrometer operates at a high vacuum, while decompositions are most often carried out at 1 atm.

The methods for the use of mass spectrometers in conjunction with thermal analysis units are as follows.

1. Collect a sample and take it to the mass spectrometer.
2. Use an interface.
3. Weigh the product gas.
4. Put the sample in the mass spectrometer.

1. Collection of Gas Samples Separately

The simplest scheme is simply to collect gas samples in glass vials and then present each vial to the mass spectrometer. If the mass spectrometer is being used for a wide variety of studies, this is the most economical approach. It has the following advantages.

1. The equipment is simple.
2. Its use is economical.
3. It is suitable for determination of kinetics. By this method the kinetics of individually evolved gas species can, in principle, be related to available surface reaction sites (not widely exploited).
4. The overall reaction can be deduced if all gas species involved are detected and evaluated.

2. Use of the Mass Spectrometer to Weight the Product Gas

In this method the gas from the reaction vessel at vacuum or at a known pressure is expanded into a large evacuated known volume. The sampling device then takes off the measured aliquot at this low acceptable pressure into the mass spectrometer. The calculation is based on the use of the Gas Laws. It has been used extensively by John Dollimore and co-workers.

A typical use of this method is a study of carbon oxidation kinetics by a temperature jump method. It should be used as a TG unit when the sample is at extremely high temperatures. This avoids complications in conventional TG equipment when used at very high temperatures.

3. Samples Placed in the Mass Spectrometer

This method has been used by Gallagher (1978) and by Price *et al.* (1980). In these units the temperature control of the sample and temperature programming is provided. The method decomposes the sample under a very high vacuum. The product gases have little time to decompose, so side reactions are eliminated. It should be noted that because of these unconventional features the results should not necessarily coincide with conventional data. The product gases are analyzed by mass number, but the solid residue cannot be determined *in situ*.

In the time-of-flight mass spectrometer equipment described by Price *et al.* (1980) the sample is decomposed in an open tube with close "line-of-sight" access to the ion source of the mass spectrometer. The sample is subjected to a linear temperature rise, with all gases rapidly removed using a powerful diffusion pump. The pressure in the ion source is always lower than 10^{-4} nm^{-2}. The ion current at the appropriate mass number is then proportional to the rate of gas evolution. If this is the only gas, it is also proportional to the mass loss on decomposition.

D. Infrared Spectroscopic Methods

Infrared spectroscopy methods are used for water vapor, CO, and CO_2. These are especially difficult gases to analyze accurately on a mass spectrometer. The infrared spectroscopy methods is also especially suited for on-stream

TABLE IX Mass Spectrometric Thermal Analysis

	Mass spectrometric analysis	Mass spectrometric thermal analysis	Mass spectrometer coupled to DTA or TG
Pressure control	That of mass spectrometer—No	That of mass spectrometer—No	Variable—Yes
Temperature control	Sample dependent—Some	Variable—Yes	Variable—Yes
Rate control	Sample dependent—No	Variable—Yes	Variable—Yes

analysis. It is not restricted to the above gases, of course, but as noted there is special merit in applying the method to them.

In an infrared radiometer, the output from an infrared source is split to pass through the sample and reference effluent streams, and the radiation intensities are then compared. The differential output is a measure of the extra radiation absorbed by the evolved gas in the sample stream. Such a system is not specific to one gas, and the response factor varies with the absorption coefficients of the component gases.

Infrared spectroscopy may be applied to EGA in two ways.

1. *Continuous monitoring of specific products*. This is a development of the infrared radiometer, in which the detector is sensitized to only one component, either by including filters (a nondispersion analyzer) or by operating at a single wavelength (a dispersion analyzer). A series of such analyzers can be used, each detecting one gas. Commercial units are available to detect CO_2, CO, and H_2O. Such methods can be used to establish kinetic parameters associated with the evolution of these gases.
2. *Special analysis of evolved gases*. When the composition or identity of the components in a gas stream needs to be established, then a spectral analysis of evolved gases is required. Obtaining such data is relatively easy but the method is noncontinuous with respect to time.

E. Gas Chromatographic Methods

There are numerous descriptions of the coupling of gas chromatography with thermal analysis units. Chiu (1968, 1970) describes a TG unit coupled with a gas chromatograph. In the use of gas chromatography the gas detector device used has to be suitable to match the gases evolved; otherwise they will escape detection. The three most commonly employed are thermal conductivity detectors, gas density detectors, and ionization detectors. The restriction on the employment of gas chromatography coupled with thermal analysis units is that the analysis is intermittent and not continuous.

V. THERMOMECHANICAL METHODS

A. Introduction

Thermodilatometry (TDA) has been defined as a technique in which a dimension of a substance under a negligible load is measured as a function of temperature while the substance is subjected to a controlled temperature program. The record is termed the thermodilatometric curve. The dimension should be plotted on the ordinate, increasing upward, and T or t on the abscissa, increasing from left to right. Linear TDA is then the measurement of one dimension of a solid “form” against altering temperature. Volume TDA measures the change in volume of the solid or other phase versus the temperature program to which it is subjected. It should be noted that this means that any measure of change of density versus temperature of treatment falls within the definition of TDA.

Thermomechanical analysis (TMA) is a technique in which the deformation of a substance under a nonoscillatory load is measured as a function of temperature while the substance is subjected to a controlled temperature program. The mode as determined by the type of stress applied (compression, tension, flexure, or torsion), which should always be stated.

As in other cases, when reporting TMA the temperature environment of the substance should be clearly stated. The type of deformation (tensile, torsional, bending, etc.) and the dimensions, geometry, and materials of the loading elements should be noted. The ordinate scale should also be identified in specific terms where possible. For static procedures, increasing expansion, elongation, or extension and torsional displacement should be plotted upward. Increased penetration or deformation in flexure should be plotted downward. For dynamic mechanical procedures, the relative modulus and/or mechanical loss should be plotted upward.

In dealing with change in the dimensions of a sample with temperature the measurements relate to the deformation and strength of the sample. Expansion in a solid or a liquid is indicative of a decrease in surface energy. Thus, the expansion of a solid in an adsorption process may be interpreted as a decrease in surface energy.

A practical distinction between TMA and TDA is simply that in TMA some kind of stress or load is applied to the test material, while in TDA no load or stress is required. The same basic equipment may be used. In commercial equipment the temperature may go as low as -170°C or as high as 1000°C.

B. Apparatus

A wide variety of equipment is available, which includes the measurement of various properties. The list given here, with brief descriptions, is illustrative rather than comprehensive. Logically volume changes and measurement of density should be discussed under this heading. However, most commercial TMA units note only a single dimensional change on fabricated units (it is difficult to make TMA measurements on powders). The measurement of density involves techniques which would be difficult to

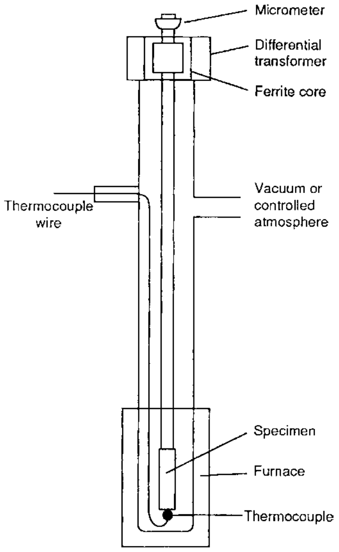


FIGURE 12 Schematic diagram of a dilatometer unit operating in a thermal analysis mode.

perform as a function of temperature. Again, logically the term *TMA* covers mechanical effects other than change in a linear dimension. Two techniques are often described: (1) TDA and (2) TMA. A third technique, dynamic mechanical analysis (DMA), is discussed in Section VI.

1. The Thermodilatometry Apparatus

A simple TDA unit is shown in Fig. 12. This simply measures expansion or contraction of a sample under test with temperature or time. The rod used in this equipment is made of a material suitable for use in the temperature range required: up to 1000°C, fused silica; up to 1200°C, porcelain; up to 1800°C, sapphire; and up to 2000°C, graphite, molybdenum, or tungsten. In the latter case a reducing or an inert gas (argon) must be used.

The coefficient of expansion against temperature should be plotted. In the simple equipment noted above, an extension rod can be used. Alternative methods of measuring the extension (or contraction) are as follows.

1. A precursor micrometer.
2. A cathetometer.
3. An interferometer.
4. From X-ray diffraction measurements.

2. The Thermomechanical Apparatus

Thermomechanical equipment represents generally an extension of the simple dilatometer principle, in which expansion or contraction under a load or a deforming stress is applied to a sample. In practice, TDA finds extensive application to ceramics and formed inorganic materials, while TMA finds its most useful role in application to polymeric systems. Various TMA probe configurations available commercially are shown in Fig. 13.

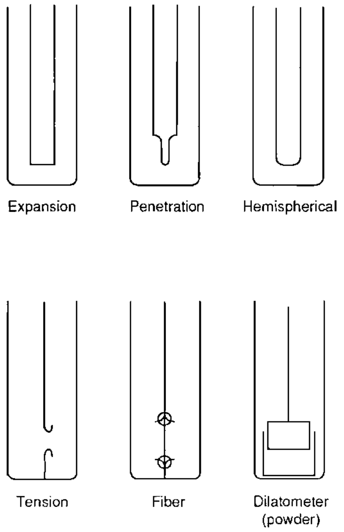


FIGURE 13 TMA probe configurations.

VI. MODULATION TECHNIQUES

It has become increasingly popular to study modulation techniques. In such techniques there is a modulation of certain experimental parameters. Two methods are discussed here: dynamic mechanical analysis (DMA) and modulated DSC.

A. Dynamic Mechanical Analysis

In dynamic thermomechanometry the dynamic modulus and/or damping of a substance under an oscillatory load is measured as a function of temperature while the substance is subjected to a controlled temperature. The frequency response is then studied at various temperatures. Torsional braid analysis is a particular case of dynamic thermomechanometry in which the substance is supported on a braid. These are all sophisticated versions of thermomechanical methods. The word *dynamic* here, as noted above, means oscillatory and this term can be used as an alternative to modulation. In DMA the sample is oscillated at its resonant frequency, and an amount of energy equal to that lost by the sample is added in each cycle to keep the sample oscillation at a constant amplitude.

This technique measures the ability of materials to store dissipate mechanical energy on deformation. If a material, for example, is deformed and then released, a portion of the stored deformation energy will be returned at a rate which is a fundamental property of the material. That is, the sample goes into damped oscillation. For an ideal elastic

material with a high Q (quality factor), the energy incorporated into oscillation will be equal to that introduced by deformation, with the frequency of the resultant oscillation being a function of the modulus (stiffness) of the material. Most real materials, however, do not exhibit ideal elastic behavior but, rather, exhibit viscoelastic behavior in which a portion of the deformation energy is dissipated in other forms such as heat. The greater this tendency for energy dissipation, the larger the damping of this deformation-induced oscillation. On the other hand, if this dissipated energy is continually made up (by an in-phase drive signal applied to the system), the sample will stay in continuous natural frequency (compound resonance) oscillation.

The two properties measured by DMA are the resonant frequency and energy dissipation, and these can be measured over a wide range of temperatures and moduli.

The resonant frequency obtained is related to the Young's modulus of the sample via the following equation:

$$E = \left[\frac{4\pi^2 f^2(J-K)}{2w((L/2)+D)^2}\right]\left(\frac{L}{T}\right)^3,$$

where

E = Young's modulus,
f = DMA frequency,
J = moment of inertia of the arm,
K = spring constant of the pivot,
D = clamping distance,
w = sample width,
T = sample thickness,
L = sample length.

The energy dissipation obtained is related to properties such as impact resistance, brittleness, and noise abatement.

B. Temperature Modulated Differential Scanning Calorimetry (MDSC)

In this technique the linear temperature regime impassed in DSC is replaced by a sinusordal temperature modulation superimposed on a linear (constant) heating profile.

The program sample temperature ($T(t)$) in normal DSC is given by

$$T(t) = T_o + \beta t,$$

where T_o (K), β (K/min), and t (min) denote the starting temperature, linear constant heating or cooling rate, and time, respectively.

A sinusoidal modulation would then be represented by

$$T(t) = T_o + \beta t + A_T \sin \mathrm{wt},$$

where A_T (±K) denotes the amplitude of the temperature modulation, w (s^{-1}) is the modulation frequency, and

$$w = \frac{2\pi}{p},$$

where p (s) is the modulation period. By using those points which lie on the linear temperature profile, a "normal" signal can be obtained together with the modulated signal. In both sets of experimental data the total heat flow at any point is given by

$$\frac{dQ}{dt} = Cp\beta + f(T_1 t),$$

where Q (J) denotes the heat, t (s) the time, Cp (J/K) the sample heat capacity, and $f(T_1 t)$ the heat flow from kinetic processes which are dependent on both temperature and time.

The linear temperature programmed DSC measures only the total heat flow. However, the sinusoidal heating profile gives the heat capacity data corresponding to the rate of temperature change. The heat capacity component of the total heat flow, $Cp\beta$, is called the reversing heat flow and the kinetic component, $f(T_1 t)$, is called the nonreversing heat flow. In one experiment the heat capacity can then be calculated using a discrete Fourier transformation by the relationship

$$Cp = K\frac{(Q_{\mathrm{amp}})}{(T_{\mathrm{amp}})}\frac{(P)}{(2\pi)},$$

where k denotes the heat capacity constant, Q_{amp} the heat flow amplitude, and T_{amp} the temperature amplitude.

The reversing component of the total heat flow signal ($Cp\beta$) allows the nonreversing component to be calculated using the relationship

nonreversing heat flow = total heat flow − reversing heat flow.

The technique finds advantages over normal DSC in calculation of the heat capacity and in determination of the glass transition point.

VII. SIMULTANEOUS TECHNIQUES

A. Introduction

It is becoming a common practice to apply two (or more) techniques of thermal analysis to the same sample at the same time. The most common application is that of TG and DTA. However, it should be noted that EGA is rarely used as a separate technique and is most often used in a simultaneous combination. In writing, the names of simultaneous techniques should be separated by the use of

the word *and* when used in full and by a hyphen when abbreviated acceptably, e.g., simultaneous TG-DTA. Unless contrary to established practice, all abbreviations should be written in capital letters.

The other practice of thermal analysis in which more than one technique is used is *coupled simultaneous techniques*. This term covers the application of two or more techniques to the same sample when the two instruments involved are connected by an interface, for example, simultaneous DTA and mass spectrometry. The term *interface* refers to a specific piece of equipment that enables the two instruments to be joined together. In coupled simultaneous techniques as in discountinuous simultaneous techniques, the first technique mentioned refers to the first in time measurement. Thus, when a DTA instrument and a mass spectrometer are used together, then DTA-MS is correct.

The name *discountinuous simultaneous techniques* is also found, and this term covers the application of coupled techniques to the same sample when sampling for the second technique is discontinuous, for example, discontinuous simultaneous DTA and gas chromatography when selected portions of evolved volatiles are "sampled" on the gas chromatography unit. Another case of a simultaneous technique arises when two samples are subjected in the same furnace to two different thermal analysis techniques, for example, DTA and TG.

B. Advantages and Disadvantages of Simultaneous Techniques

The advantage of using simultaneous techniques lies in the nature of the samples being investigated. They are a reflection of the fact that in certain materials there is a distinct probability of different samples showing different signals. One can cite impure samples where the signal may vary with the nature and amount of the impurity. In other samples studied by simultaneous techniques such as pharmaceutical stearic acid or magnesium starate (known to contain other carboxylic acids), coal samples, and other natural products, the sample may vary from one region to another. Simultaneous tests on one sample may then represent a real advantage.

It must not be imagined that the combination of techniques should always produce the same results shown by the separate techniques, and this is particularly true for DTA and TG. Consider first the mass of material studied. Obviously the mass of the substance used in a TG run should be sufficient to note the change in mass accurately. However, the ideal mass of a sample used in DTA to achieve the best results may be much smaller than in TG. The sample holders in many DTA experiments are sealed. The use of such a technique in DTA-TG would be to inhibit or prevent mass loss. Crucibles ideal for TG may not be ideal for DTA. Finally, the heating rate may be cited: to obtain good DTA experiments the heating rate should be reasonably fast; otherwise the peaks are broad and shallow. A slow heating rate, however, often gives the best results in TG experiments.

C. New Studies Made Possible by Simultaneous Techniques

There are cases where simultaneous techniques can lead to interpretations not possible using convention techniques separately. One such study is evaporation. In conventional DTA, with the sample in a closed crucible, such a study is impracticable. However, with the open crucible, evaporation in a TG-DTA unit shows a distinctively shaped endothermic peak; the TG for one-component systems demonstrates a zero-order process; and, using the Langmuir equation for evaporation, Price and Hawkins show that a vapor-pressure curve (partial pressure plotted against temperature) can be obtained. Calibration for temperature also becomes easy and more reliable. Combination of techniques in simultaneous units also allows interpretations of complex degradations to be studied. Thus it can be established that $Mg(NO_3)_2 \cdot 6H_2O$ melts at 90°C and then water is lost from the system, leaving an anhydrous solid salt at 350°C. This melts at 390°C and decomposes to the oxide at 600°C. DTA shows that all the processes are endothermic; gas analysis shows the loss of water and identifies the gases evolved in the final dissociation to the oxide. Hot-stage microscopy has been used to establish the melting processes.

VIII. THE ROLE OF THE COMPUTER

In some of the early designs for dedicated computers the computer was completely dedicated to the equipment and operated only with programs provided by the instrumentation company. To appreciate the role of the computer, however, it must be noted that it should play three roles.

1. To control the programs that alter the temperature regime.
2. To record the property being measured and demonstrate its variation with the temperature regime.
3. To interface this dedicated computer with computer programs or transfer the data to other computers and allow the operator in charge of the study to manipulate the data and derive appropriate parameters.

In the case of determination of kinetic data from raw experimental data, this usually means feeding such data (e.g., mass and temperature) into a spreadsheet and calculating the preexponential term and the activation energy in the Arrhenius equation.

Most modern dedicated computers fulfill these three roles and allow the operator to manipulate the basic experimental data by using his own spreadsheets or computer programs. It must be noted that instrument manufacturers supply their own programs to make sure that the use of the equipment in industry is made as easy as possible.

IX. APPLICATION OF THERMAL ANALYSIS TECHNIQUES

A. Introduction

The application of thermal analysis techniques is based on either thermodynamic considerations or the kinetics of change. It is not just the fact that the temperature is being changed that makes a choice of this kind necessary, but the kind of systems investigated. Some macromolecules are so big that phase changes which, in simple systems, would show only thermodynamic (equilibrium) features exhibit a kinetic factor. In the preceding survey certain techniques have not been reported or have not been described in detail. Only the main techniques available from more than one commercial manufacturer have been reported in some detail.

B. Thermodynamic Considerations

In a condensed system, if a phase change occurs in the heating mode where one phase is stable over a definite temperature range and the other phase is stable over, say, a higher temperature range, then the process of change at the transition temperature will be endothermic and reversible on the cooling mode with an exothermic character. This can be represented as

$$\text{A}(1) \xrightarrow[\text{heating}]{} \text{A}(2) \text{ endothermic}$$

$$\text{A}(2) \xrightarrow[\text{cooling}]{} \text{A}(1) \text{ exothermic}$$

This is true where both phase (1) and phase (2) are solids or phase (1) is solid and phase (2) is liquid. In the phase change from a liquid to a gas, the process is reversible, but because the gas is usually lost from the system, it is probably best to follow the process at different partial pressures of the vapor (p) when the relationship

$$\ln p = \frac{-\Delta H}{RT} - \text{constant} \tag{6}$$

holds, where R is the gas constant, T is the temperature of the phase change (K), and ΔH is the enthalpy of fusion. The above phase transitions can all be investigated by DTA or DSC.

If an amorphous (energy-rich) solid phase undergoes transition, however, then a stable crystal may result. The amorphorus or glassy phase is metastable, and the process is exothermic and irreversible. This can be represented as

Metastable	→	Stable form
(energy rich)		(crystalline)
Exothermic		and irreversible

Such changes are shown in both inorganic and organic systems.

The above are all examples of transitions which are of first order. However, second-order transitions can also be followed and determine the position of the baseline in DSC. This leads to the calculation of heat capacity from DSC measurements and is the basis for the establishment of the glass transition temperature (T_g) in polymers. A further extension based on thermodynamic factors is purity determination based on the lowering of the freezing point caused by the presence of impurities.

In chemical reactions certain features may be noted. Thus, carbonate decompositions are endothermic, and loss of water is also endothermic. Obviously, such processes involving weight loss can be studied using both TG and DTA/DSC. It may also be advantageous to study such processes under partial pressure (P) of the gaseous products when one may expect the relationship

$$\ln p = \frac{-\Delta H}{RT} + \text{constant}, \tag{7}$$

similar to the liquid–gas phase change, but here ΔH is the enthalpy of the reaction.

C. Kinetic Considerations

Where kinetic factors are important in studying systems using thermal analysis techniques, then the Arrhenius parameters in theory should be able to be estimated. The Arrhenius equation can be written

$$k(T) = Ae^{-E/RT}, \tag{8}$$

where $k(T)$ is the specific reaction rate constant and is a function of temperature, A is the preexponential term, E is the energy of activation, R is the gas constant, and T is the temperature (K). In the rising temperature mode we have

$$d\alpha/dt = k(T)f(\alpha), \tag{9}$$

where α is the fraction decomposed at time t, $d\alpha/dt$ is the rate of the reaction, $k(T)$ is the specific rate constant at

temperature T, and $f(\alpha)$ is a score function of α describing the progress of the reaction. The temperature regime imposed on the system can be represented by

$$T = T_0 + bt, \tag{10}$$

where T is the temperature at time t, T_0 is the starting temperature, and b is the heating rate. Combination of all these equations, noting that in linear temperature programming $b = dT/dt$, gives

$$k(T) = \frac{(d\alpha/dT)b}{f(\alpha)}, \tag{11}$$

thus allowing k to be plotted against $1/T$ if $f(\alpha)$ can be identified, whence from the relationship

$$\ln k = \ln A - (E/RT),$$

both A and E can be calculated. The difficulty in obtaining $d\alpha/dT$ with sufficient precision has led to extensive studies into the use of an integral method. This entails the evaluation of

$$\int e^{-E/RT}\, dT,$$

and this is impossible to do analytically. Numerous methods of overcoming this difficulty have been suggested, usually based on the numerical methods of integrations or alternative similar expressions that can be analytically evaluated. The solid state is the only phase for which one has to state a prehistory, and an article by Flynn (1981) deals with this aspect. Related to these studies are the prediction of shelf life, especially of pharmaceuticals, and the proximate analysis of coal.

SEE ALSO THE FOLLOWING ARTICLES

Critical Data in Physics and Chemistry • Heat Transfer • Kinetics (Chemistry) • Physical Chemistry • Thermodynamics • Thermometry

BIBLIOGRAPHY

Adamson, A. A. (1986). "A Textbook of Physical Chemistry," 3rd. ed., p. 401, Academic Press, Orlando, FL.

Atkins, P. W. (1982). "Physical Chemistry," 2nd ed., p. 298, Freeman, San Francisco.

Blaine, R. L., and Fair, P. G. (1983). *Thermochim. Acta* **67,** 233.

Boersma, L. (1955). *J. Am. Ceram. Soc.* **38,** 281.

Brown, M. E., Dollimore, D., and Galwey, A. K. (1980). Reactions in solid state. *In* "Comprehensive Chemical Kinetics" (C. H. Bamford and C. F. H. Tipper, eds.), Vol. 22, p. 99, Elsevier Science.

Charsley, E. L., Rumsey, J. A., and Warrington, S. B. (1984). *Anal. Proc.* XX, 5.

Chiu, J. (1968). *Anal. Chem.* **40,** 1516.

Chiu, J. (1970). *Thermochim. Acta* **1,** 231.

Cunningham, A. D., and Wilburn, F. W. (1970). "Differential Thermal Analysis" (R. C. Mackenzie, ed.), Vol. 1, p. 31, Academic Press, London.

David, D. J. (1964). *Anal. Chem.* **36,** 2162.

Dollimore, D., and Reading, M. (1993). Application of thermal analysis to kinetic evaluation of thermal decomposition. *In* "Treatise on Analytical Chemistry, Part I. Thermal Methods," 2nd ed., (J. D. Winefordner, D. Dollimore, and J. Dunn, eds.), Vol. 13, pp. 1–61, John Wiley & Sons, New York.

Dollimore, J., Freedman, B. H., and Quinn, D. F. (1970). *Carbon* **8,** 587.

Dunn, J. G., and Sharp, J. H. (1993). Thermogravimetry. *In* "Treatise on Analytical Chemistry, Part I. Thermal Methods," 2nd ed. (J. D. Winefordner, D. Dollimore, and J. Dunn, eds.), Vol. 13, pp. 127–266.

Flynn, J. H. (1981). *In* "Thermal Analysis in Polymer Characterization" (E. A. Turi, ed.), p. 43, Heyden, Philadelphia.

Gallagher, P. K. (1978). *Thermochim. Acta* **26,** 175.

Galwey, A. K., and Brown, M. E. (1999). "Thermal Decomposition of Ionic Solids," p. 597, Elsevier, Amsterdam.

Grimshaw, R. W., Heaton, E., and Roberts, A. L. (1945). *Trans. Br. Ceram. Soc.* **44,** 76.

Hatakeyama, T., and Quinn, F. X. (1999). "Thermal Analysis," 2nd ed., p. 180, Wiley, Chichester.

Hill, J. O. (1991). "For Better Thermal Analysis," 3rd ed., ICTA.

Mackenzie, R. C., and Mitchell, B. D. (1970). "Differential Thermal Analysis" (R. C. Mackenzie, ed.), Vol. 1, p. 63, Academic Press, London.

Marti, E. (1972). *Thermochim. Acta* **5,** 173.

McGhie, A. R. (1983). *Anal. Chem.* **55,** 987.

McGhie, A. R., Chiv, J., Fair, P. G., and Blaine, R. L. (1983). *Thermochim. Acta* **67,** 241.

Norem, S. D., O'Neill, M. J., and Gray, A. A. (1970). *Thermochim. Acta* **1,** 29.

O'Neill, M. J. (1964). *Anal. Chem.* **36,** 1233.

Price, D., Dollimore, D., Fatemi, N. J., and Whitehead, R. (1980). *Thermochim. Acta* **42,** 517.

Prout, E. G., and Tompkins, F. C. (1946). *Trans. Faraday Soc.* **43,** 482.

Radecki, A., and Wesolowski, M. (1979). *J. Therm. Anal.* **17,** 73.

Rosenvold, R. J., Dubow, J. B., and Rajeshwar, K. (1982). *Thermochim. Acta* **53,** 321.

Shoemaker, D. P., and Garland, C. W. (1967). "Experiments in Physical Chemistry," 2nd ed., p. 177, McGraw–Hill, New York.

Vold, M. J. (1949). *Anal. Chem.* **21,** 683.

Watson, E. S. M., O'Neill, M. J., and Brenner, N. (1964). *Anal. Chem.* **36,** 1233.

Tomography

Z. H. Cho

Korea Advanced Institute of Science

I. Introduction
II. Principles of Computerized Tomography
III. Areas of Application
IV. Recent Developments

GLOSSARY

Algorithm Set of well-defined rules for solving a problem in a finite number of steps.

Coincidence detection Detection method in which an event is registered only if two photons are detected within a specified, sufficiently short time interval.

Fast Fourier transform (FFT) Highly optimized Fourier transform algorithm for digital computation.

Free induction decay (FID) Nuclear magnetic resonance signal emitted by precession of transverse magnetizations after excitation.

Gradient coils Electromagnetic coils generating magnetic fields, which are superimposed on the main magnetic field to create spatial variation in the field strength.

Gray level Discrete steps between light and dark in the image.

Linear attenuation coefficient Probabilities per unit path length that the X-ray photon will be removed from the beam. This includes the effects of photoelectric absorption, Compton scatter, and pair production.

Monochromatic Refers to an electromagnetic wave with a negligibly small region of spectrum.

Nyquist sampling criterion Criterion of the allowable maximum sampling interval that can be given to regularly spaced sampled data of a signal with bandwidth B for complete determination of the signal in its original form. It corresponds numerically to $1/(2B)$.

Pixel Abbreviation for "picture element"—a basic element in digital image.

Point-spread function (PSF) Transfer function that represents the output of a system to an infinitely high amplitude point input.

Polychromatic Antonym of monochromatic.

Radionuclide Radioactive nuclei undergoing nuclear transitions that are usually accompanied by the emission of particles or electromagnetic radiation.

RF Pulse Pulse of radio-frequency (RF) energy transmitted by an RF coil in nuclear magnetic resonance computerized tomography. Its frequency ω_0 is represented as $\omega_0 = \gamma H_0$, where γ is the gyromagnetic ratio and H_0 is the main magnetic field strength. Its shape and amplitude determine the selectivity in the frequency band and the amount of spin rotation.

Scintigraphy Imaging technique that uses an Auger (or γ) camera to visualize the distribution of radioisotopes within the human body.

Spin Property of nuclei that have an odd number of neutrons and protons. Nuclei with spin have a magnetic moment and can possess the NMR property.

Spin echo Signal produced by the 90°–180° RF pulse sequence. The spin-echo signal is actually conjugate symmetric to its center if its duration is short compared with T_2. Its amplitude is determined by the T_2 of the substance, excluding the effects of field inhomogeneity.

Spin–lattice relaxation (thermal or longitudinal relaxation) Phenomenon of spins going to the thermal equilibrium state with other molecules in lattice. It is characterized by the exponential time constant T_1. Also, it determines the recovery time of the longitudinal magnetization M_z.

Spin–spin relaxation (transverse relaxation) Exchange of energy of excited nuclei with other precessing nuclei. It is characterized by the exponential time constant T_2. Also, it determines the envelope of the free induction decay in a perfectly uniform magnetic field and the amplitude of the spin echo.

Superconductive magnet Magnet that requires no electrical power once the field has been established. Superconductivity is a property of some materials that have no electrical resistance when the temperature is near absolute zero. Liquid helium is generally used to maintain a low temperature.

True coincidence Event detected in coincidence without scatter of either photon, generated from an annihilated positron, in the object.

COMPUTERIZED or computed tomography (CT) is a technique of producing cross-sectional three-dimensional images from multiple views or projection data obtained with penetrating probe radiations or by other means such as magnetic field gradients, by processing those data using a computer and mathematical image reconstruction algorithms. Major applications of the basic CT concepts are medical diagnosis, industrial nondestructive testing, and other areas of the physical sciences, such as geophysical exploration.

I. INTRODUCTION

Computerized tomography is a technique by which three-dimensional (3-D) imaging of an object is made possible. The basic data are most often obtained in the form of projection data. Since the first tomographic system, known as X-ray CT, was designed by G. N. Hounsfield in 1972, many applications of the technology, based on 3-D image reconstruction from projection, have been developed. These include X-ray CT, radionuclide (isotope) emission CT (ECT), and nuclear magnetic resonance (NMR) CT. The basic forms of projection data are obtained from focused and collimated X rays, γ rays, and annihilation photons from decaying radionuclides, transmitted and reflected ultrasound beams, appropriately excited nuclear spins under a strong magnetic field, and so on. The data obtained, usually referred to as projection data, are processed by a mathematical image reconstruction algorithm using a digital computer to form an image or a set of images, each representing a slice or several slices of an object. Although direct Fourier transform image reconstruction is employed in some cases to obtain an image or a set of images, as in NMR CT, most types of image reconstruction employ some forms of projection reconstruction. Those mathematical techniques known as 3-D image reconstruction form the basis of all the CTs discussed in this article. Finally, mathematically formed or reconstructed images are displayed on a television screen or photographed by a camera attached to the system.

From the physics and engineering points of view, both 3-D image reconstruction and CT are new concepts and tools. For the first time in history, human beings are capable of visualizing the inner structures of an object noninvasively. At present, the most widely used CT is X-ray CT; it is estimated that more than 5000 X-ray CT units, each costing as much as $1 million, are in operation throughout the world. In the area of ECT, two types are under active development: single photon emission CT (SPECT) and positron emission tomography (PET). Although ECT is similar to X-ray CT, it differs in functional form; that is, X-ray CT is usually capable of visualizing anatomical details, while ECT is capable of visualizing the functional or metabolic behaviors of an object. Ultrasound CT is similar to X-ray CT and is capable of visualizing cross-sectional images. The development of ultrasound CT has been relatively slow, due mainly to the difficulties inherent in the basic properties of ultrasound, such as scattering and diffraction. The most recent and probably most exciting development in the field of CT is NMR CT. It is now capable of visualizing in three dimensions the distribution of several nuclei, such as the proton, the sodium nucleus, and the phosphorus nucleus, and thereby performing chemical imaging. As the resolution and sensitivity of NMR CT improve rapidly, it is becoming one of the most promising medical diagnostic imaging modalities.

Computerized tomography is believed to be the most important development in diagnostic imaging history since the discovery of X rays in 1895 by W. K. Roentgen. Applications of CT concepts are rapidly expanding from medical imaging to many branches of the physical sciences. The principles of X-ray CT are now being applied to many diverse areas, ranging from the examination of defects in nuclear reactor cores to the rapid inspection of automobile tires on the production line. The principles

of NMR CT are being applied in the area of fluid dynamics to investigate structural changes in flow. This was not feasible in the past. Future applications of CT are unlimited; they range from the physical exploration of oil to the study of mummies excavated from ancient pyramids. Through CT, human vision has expanded beyond two dimensions to three. Until recently, it was believed that 3-D imaging was impossible, considering the fundamental nature of the physical world.

This chapter discusses in detail the mathematical foundation of CT technology, specifically image reconstruction algorithms, computers and peripherals used in CT, display systems for CT, and the physics of the data collection mechanism of various CT systems (X-ray CT, ECT, and NMR CT).

II. PRINCIPLES OF COMPUTERIZED TOMOGRAPHY

A. Mathematics and Algorithms for Image Reconstruction

Several mathematical image reconstruction algorithms used for CT have been developed since the basic concept of CT was established. These algorithms are based primarily on mathematical theory as described in detail in this section, which presents a mathematical basis for the reconstruction algorithms currently used in many CT modalities, including X-ray CT, ECT, and NMR CT. According to data processing methods, algorithms can be classified into projection reconstruction, iterative method, and direct Fourier imaging. A classification of these three general methods is given in Table I.

TABLE I Image Reconstruction Algorithms

Class	Specific method
Projection reconstruction	2-D Projection reconstruction
	Filtered backprojection (FB)
	Parallel beam mode
	Fan beam mode
	Backprojection filtering (BF)
	3-D Projection reconstruction
	True 3-D reconstruction (TTR)
	Generalized TTR
	Planar-integral projection reconstruction (PPR)
Iterative method	Algebraic reconstruction technique (ART)
	Maximum likelihood reconstruction (MLR)
Fourier reconstruction	Direct Fourier reconstruction (DFR)
	Direct Fourier imaging in NMR (DFI)

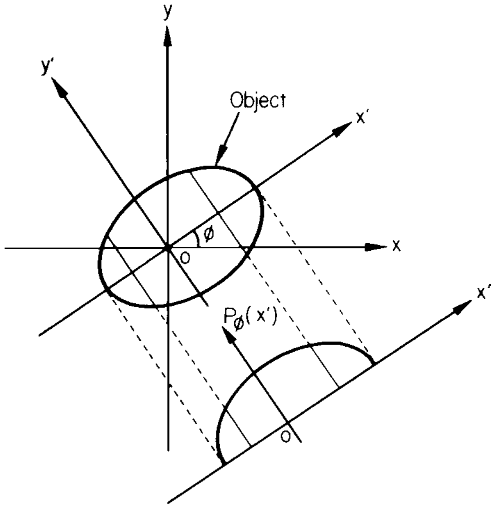


FIGURE 1 Geometry for 2-D parallel beam projection.

1. Projection Reconstruction

Since 1970, active research on image reconstruction from projection has been initiated mainly through the development of CT. This projection reconstruction is now applied to many areas of science. The most important area of its application has probably been CT; projection reconstruction has remained the basic algorithm for CT since the inception of X-ray CT in 1972.

The simplest form of projection data is illustrated in Fig. 1. The line integrals of a physical object are estimated along straight lines. Each line integral, in practice, represents a physical property in a strip with a finite width. Therefore, a set of line-integral data is obtained at each view. By repeatedly assessing the sets of data at different views, that is, around 180° or 360° with a specified angular interval $\Delta\theta$, a complete projection data set can be obtained. Each set of estimated line integrals is often called a projection or a line-integral projection. The collection of all these estimated line-integral sets around 180° or 360° is referred to as line-integral projection data or simply projections. Similarly, planar-integral projection data can also be obtained; they represent the collection of all integral values of a physical object along planes.

From a set of measured projection data or projections, an image can be formed through the use of appropriate processing algorithms. Image reconstruction from projections is the process of producing an image of a 2-D or a 3-D distribution of some physical property from the estimates of its line (or plane) integrals along a finite number of lines (or planes) of known locations. For the reconstruction of the images from projections, several algorithms can be used depending on the imaging modalities. Theoretical aspects and characteristics of these algorithms are discussed in the following subsections. Fourier transforms,

which constitute the main body of image reconstruction in general, are also briefly explained in the following. Fourier transforms are basically part of a conversion process that converts spatial domain data to spatial frequency domain data. They are defined as

$$F(\omega) = \mathscr{F}_1[f(x)] = \int_{-\infty}^{\infty} f(x)\exp(-i\omega x)\,dx \quad (1)$$

$$f(x) = \mathscr{F}_1^{-1}[F(\omega)] = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)\exp(i\omega x)\,d\omega, \quad (2)$$

where $i = \sqrt{-1}$ and $\mathscr{F}_1[\cdot]$ and $\mathscr{F}_1^{-1}[\cdot]$ are the 1-D forward and inverse Fourier transform operators, respectively.

a. Two-dimensional projection reconstruction.

i. Filtered backprojection algorithm. The filtered backprojection (FB) or convolution backprojection algorithm is the most popular and most frequently used reconstruction method so far employed in CT, with the exception of NMR CT. For the mathematical formulation of this FB algorithm, there are two basic forms in existence: the parallel beam and fan beam modes.

Parallel Beam Mode. Let us assume an object distribution function $f(x, y)$ represented by the Cartesian coordinates (x, y), with rotated coordinates expressed as (x', y'). The basic data to be used in the reconstruction are projections that represent sets of line integrals of an object in various directions. The projection data $p_\phi(x')$ shown in Fig. 1 is a set of line integrals taken along the y' direction, that is,

$$p_\phi(x') = \int_{-\infty}^{\infty} f(x', y')\,dy' = \iint_{-\infty}^{\infty} f(x, y)\delta(x\cos\phi + y\sin\phi - x')\,dx\,dy, \quad (3)$$

where

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}. \quad (4)$$

The Fourier transform of projection data $P_\phi(\omega)$ can be related to the projection data $p_\phi(x')$ as

$$P_\phi(\omega) = \int_{-\infty}^{\infty} p_\phi(x')\exp(-i\omega x')\,dx' = F(\omega_x, \omega_y)|_\phi = F(\omega, \phi), \quad (5)$$

where $F(\omega_x, \omega_y)$ is the 2-D Fourier transform of $f(x, y)$, $\omega_x = \omega\cos\phi$, and $\omega_y = \omega\sin\phi$ when (ω_x, ω_y) and (ω, ϕ) represent the Cartesian and polar coordinates of (x, y) in the spatial frequency domain, respectively.

Equation (5) states that a 1-D Fourier transform of projection data at a given angle ϕ represents the 2-D Fourier transform values of the object function $f(x, y)$ in the spatial frequency domain along the radial frequency with a given angle ϕ. This is the projection theorem that plays a central role in 2-D image reconstruction. From this theorem, it can easily be shown that the object function $f(x, y)$ can be recovered as

$$f(x, y) = \frac{1}{2\pi}\int_0^{\pi} d\phi\left[\int_{-\infty}^{\infty} dx'\, p_\phi(x')h(x\cos\phi + y\sin\phi - x')\right], \quad (6)$$

where

$$h(x') = \mathscr{F}_1^{-1}[|\omega|].$$

The convolution kernel $h(x')$ in Eq. (6) is an inverse Fourier transform of $|\omega|$, but its exact form is not realizable in practice. Therefore, several modified filter functions have been suggested. The selection of a particular filter function will affect the characteristics of the reconstructed image, that is, the desired image resolution and contrast.

Fan Beam Mode. Although the parallel beam reconstruction algorithm has been a basic tool for image reconstruction, the fan beam reconstruction algorithm is nevertheless widely used. For example, it is used in X-ray CT, due mainly to the fact that the basic data collection procedure involved is the fan beam mode. In addition, image reconstruction that utilizes the fan beam algorithm often provides better resolution with the same amount of sampled data than the parallel beam algorithm because of improved sampling at the central region.

If the fan beam projection data set is represented as $p_{\alpha_1}(\beta_1)$, where β_1 and α_1 represent the detector position and rotation angle of the beam, respectively (Fig. 2), then the relation between parallel and fan beam projection data with coordinates (x', ϕ) and (α_1, β_1) can be represented as

$$\begin{aligned} x' &= R_\mathrm{d}\sin\beta_1 \\ \phi &= \alpha_1 + \beta_1, \end{aligned} \quad (7)$$

where R_d is the distance between the center point and the apex of the fan. Through the use of Eqs. (6) and (7) and the rotation angle $\phi = 0 \sim 2\pi$, the fan beam analogy can be derived as

$$f(x, y) = \frac{1}{4\pi}\int_0^{2\pi} d\alpha_1 \int_{-\beta_{1m}}^{\beta_{1m}} d\beta_1\, p_{\alpha_1}(\beta_1)J(\beta_1)g \times (\beta_1' - \beta_1)\frac{1}{V_1^2}, \quad (8)$$

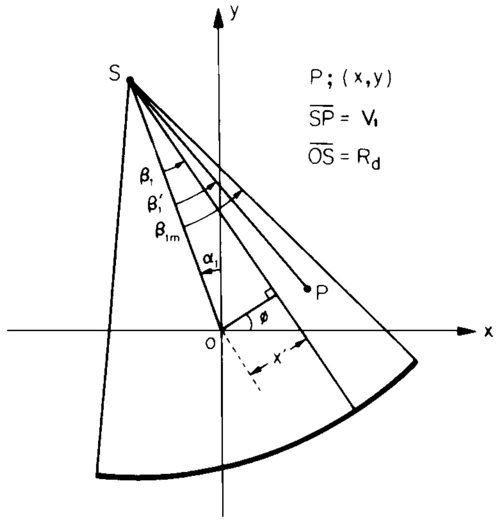


FIGURE 2 Geometry for 2-D fan beam projection.

where

$$J(\beta_1) = \left| \frac{\partial(x', \phi)}{\partial(\beta_1, \alpha_1)} \right| = R_{\rm d} \cos \beta_1$$

and

$$g(\beta_1) = \left(\frac{\beta_1}{\sin \beta_1} \right)^2 h(\beta_1). \tag{9}$$

Among the variables in the above equations, V_1 is illustrated in Fig. 2 and $h(\beta_1)$ is the same one given in Eq. (6).

The fan beam reconstruction algorithm given in Eq. (8) consists of three parts: (1) weighting the projection data with $J(\beta_1)$, (2) convolution with $g(\beta_1)$, and (3) weighted backprojection with weight $1/V_1^2$.

ii. Backprojection filtering algorithm. As mentioned previously, the FB algorithm provides a high-quality image and computational efficiency. It is by far the most popular algorithm in CT. The use of this algorithm is limited, however, to the equisampled parallel or the fan beam mode of straight line-integral projection data. Therefore, the processes of rebinning and interpolation are usually required for arbitrarily oriented rays and for more general cases. It would be difficult to apply this algorithm, for example, to line-integral data along curved lines.

An alternative, backprojection filtering (BF) algorithm, has been proposed to overcome these types of problems, and this algorithm is more general as long as the point-spread function (PSF) of a backprojected image, which can be obtained from Eq. (6) without convolution kernel, follows $1/r$ characteristics, where r is the distance from the point source. Although the method appears to be attractive and general, it is rarely used because the resultant images are usually of poorer quality than those obtained through the FB algorithm. The main reasons for poor image quality are as follows:

1. Although the projection data from the finite-size object are limited in the spatial extension, the backprojected image is unlimited. Truncation of the 2-D backprojected image data for the digital processing will therefore result in image degradation.
2. The 2-D filter function to be used in the rectangular Cartesian coordinates has slope discontinuities at the boundary (near cutoff frequency), resulting in a ring artifact. By the incorporation of proper digital signal processing techniques, however, these problems can easily be resolved.

In this section, we present a BF algorithm with which high-quality images equal to the FB images can be obtained. If the PSF of the backprojected image is $1/r$, then the relation between the object density function $f(x, y)$ and the backprojected or blurred image $b(x, y)$ is

$$b(x, y) = f(x, y) ** (1/r), \tag{10}$$

where r is the distance from the source point at origin and $**$ represents the 2-D convolution. From Eq. (10), the object function $f(x, y)$ can be obtained through the inverse Fourier transform operation,

$$f(x, y) = \mathcal{F}_2^{-1}[\omega B(\omega_x, \omega_y)], \tag{11}$$

where $B = \mathcal{F}_2[b]$ and ω is the radial spatial frequency.

The computation time required in the convolution operation in the space domain is generally longer than the processing time required in the spatial frequency domain. Let us now consider Eq. (11) to be the basic form of the BF algorithm. In the implementation of this algorithm, two previously mentioned aspects should be considered: the size and form of the backprojected image and the shape of the filter function.

In conventional 2-D image processing, an image of matrix size $N \times N$ is expanded to a $2N \times 2N$ data format, in which the outsides of the $N \times N$ center array are filled with zeros to avoid the aliasing effect arising in circular convolution. A similar procedure, however, cannot be applied to the backprojected data array, because the backprojection image data are not confined to $N \times N$. Therefore, the truncation of the backprojected image data will result in a severe truncation artifact. To avoid this artifact, the image field is expanded twice; that is full $2N \times 2N$ backprojected image data are taken instead of only the $N \times N$ array from the backprojected image. The use of these full data reduces the artifact significantly.

The selection and formulation of the 2-D filter function are also important factors in determining the image quality in 2-D BF image reconstruction because the slope discontinuity at the cutoff frequency causes ring artifact. The overall reconstruction procedures of the BF algorithm are as follows:

1. Backprojection to get $2N \times 2N$ blurred image
2. Two-dimensional fast Fourier transform of the full $2N \times 2N$ data and filtering
3. Two-dimensional inverse fast Fourier transform and adoption of $N \times N$ image in the central region
4. Normalization of the image with precalculated reference data, such as the uniform disk image

b. Three-dimensional projection reconstruction.

i. True three-dimensional reconstruction with line-integral data. In the conventional reconstruction algorithm, projections of an object are taken transaxially and are used to reconstruct the object slice by slice. The stacked slices constitute the 3-D volume image of the object.

For example, in ECT, in which radionuclides are injected into the patient, the photons are emitted isotropically in 4π directions. This slice-stacking method for a volume object does not make full use of the photons because it captures photons emitted only in the direction perpendicular to the long axis of the body, constituting only a small fraction of the total emitted photons. All the emitted photons, especially in ECT, should be collected through the use of complete spherical geometry. For image reconstruction with the data collected in 4π geometry, the direct true three-dimensional reconstruction (TTR) algorithm will be required for the maximum utilization of all the available photons.

Parallel Beam Mode. Let us consider a spherical geometry that completely surrounds the object, and let us assume that all the emitted photons are captured and rearranged into 2-D parallel data sets in 4π directions. The PSF of this system is $1/r^2$, and the backprojected blurred function is given by

$$b(x, y, z) = f(x, y, z) *** (1/r^2), \qquad (12)$$

where $***$ represents a 3-D convolution operator. The object function can be obtained as

$$f(x, y, z) = \mathscr{F}_3^{-1}[\rho] *** b(x, y, z), \qquad (13)$$

where $\rho = (\omega_x^2 + \omega_y^2 + \omega_z^2)^{1/2}$ and $\mathscr{F}_3^{-1}[\cdot]$ is a 3-D inverse Fourier transform operator. Equation (13) is essentially a 3-D convolution (deconvolution) of a simple backprojected (blurred) image. This 3-D convolution process, as well as the backprojection operation, however, would require an unusually long computation time.

As an alternative approach, the FB method has been proposed. The 3-D version of the projection theorem states that the Fourier transform of the 2-D projection represents the plane data passing through the center in the frequency domain in the same direction as the projected plane of the object. This statement can be written

$$\mathscr{F}_2[p_{\theta,\phi}(x', z')] = F(\omega_{x'}, \omega_{z'}; \theta, \phi), \qquad (14)$$

where $(\omega_{x'}, \omega_{z'})$ are the spatial frequency domain coordinates of (x', z') in the direction of (θ, ϕ). If the uniform planes that pass through the center are superimposed from all possible directions, in 3-D space, the overlapped density function becomes $1/\rho$. Therefore, to obtain the 3-D object function in the Fourier domain, the 2-D Fourier transform of each projection data set should be compensated with the factor ω before the superposition.

Image function can now be obtained by

$$f(x, y, z) = \frac{1}{4\pi^2} \int_0^{\pi} d\theta \sin\theta \times \int_0^{2\pi} d\phi \left\{ \mathscr{F}_2^{-1}[\omega] ** p_{\theta,\phi}(x', z') \right\}. \qquad (15)$$

Equation (15) shows a 3-D FB algorithm, which is in fact the backprojection of the filtered 2-D projection data in 3-D space. In this case, the 2-D filter function or kernel is simply $\mathscr{F}_2^{-1}[\omega]$.

ii. True three-dimensional reconstruction algorithm for generalized geometry. The ideal detector configuration for the ECT is a complete sphere. The reconstruction of such a spherically configured emission or transmission data consisting of sets of line-integral data was discussed in the preceding section. Three-dimensional images can be reconstructed by FB of the 2-D line-integral data sets in 3-D space. Practical system design, however, prohibits the construction of such a configuration when one considers the elongated shape of the human body as an example.

As a practical alternative, a truncated spherical configuration (Fig. 3) can be considered, and an algorithm suitable for such geometry has been developed. This 3-D image reconstruction algorithm is known as the TTR algorithm. The algorithm eventually will lead to a generalized algorithm for both 2-D slice reconstruction and 3-D volume image reconstruction of a complete spherical volume. Consider a sphere in which parts of the surface have been removed along the body axis to adapt to the shape of the human body. Although the sphere is truncated, this geometry retains spherical symmetry in the remaining spherical surface. Therefore, it also retains the possibility of reconstructing a true 3-D volume image, as explained in the following paragraphs.

The 3-D image within the reconstruction sphere, which has a radius of R_0, is obtained by summing all the images reconstructed at each slice orientation in which a complete set of 2-D projection data is provided. This sum image should be divided by the number of slice orientations reconstructed for normalization. Essentially, this is the basis of the reconstruction algorithm developed for truncated spherical geometry.

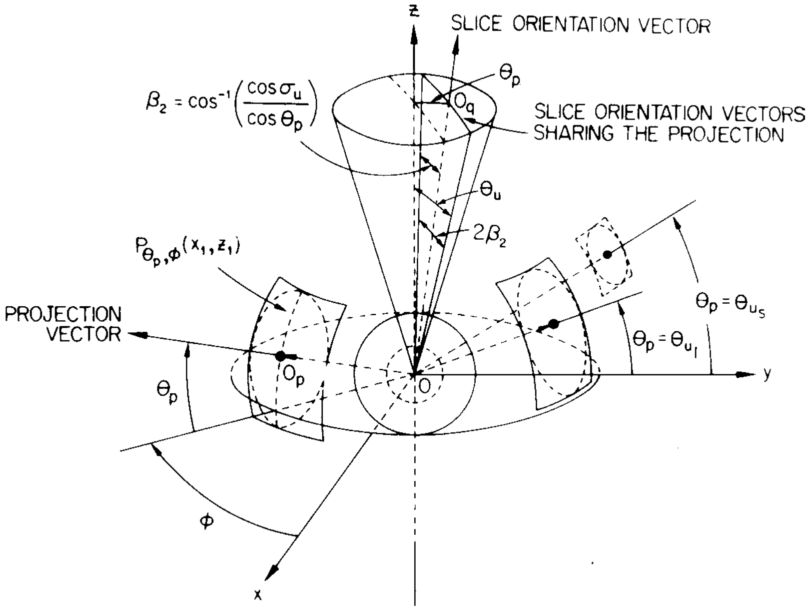


FIGURE 3 Basic geometry of the truncated spherical configuration for the development of the TTR algorithm.

For the implementation of the algorithm, all slice orientations involved in each set of projection data must be identified and the corresponding composite filter function generated. In fact, each 2-D parallel projection data set corresponding to a certain object size has a fixed number of slice orientations. This allows us to treat a certain projection data set in a unified way; that is, a 2-D projection data set can be processed with one filter function. Let us consider a 2-D projection data set projected on a direction parallel to the line OO_p in Fig. 3. Here, O_p is the center of the projection plane. As shown in the figure, these projection data are shared by slice orientations rotated around the line OO_p from $-\beta$ to β, where

$$\beta = \cos^{-1}(\cos\theta_u/\cos\theta_p). \tag{16}$$

By use of the composite filter concept, object function $f(r)$ can be obtained with the following equation:

$$f(\mathbf{r}) = \frac{\int_0^{\pi} d\phi \int_{-\theta_u}^{\theta_u} d\theta_p \cos\theta_p \int_{-\beta}^{\beta} d\alpha [p_{\theta_p,\phi}(x_1, z_1) ** h(x_1, z_1; \alpha)]}{\int_0^{2\pi} d\phi \int_{-\theta_u}^{\theta_u} d\theta_p \cos\theta_p \int_{-\beta}^{\beta} d\alpha}, \tag{17}$$

where $p_{\theta_p,\phi}(x_1, z_1)$ is the parallel projection data set at polar angle θ_p and the azimuthal angle of ϕ, and $h(x_1, z_1; \alpha)$ is the rotated form of the conventional filter function with angle α, to be applied to projection data $p_{\theta_p,\phi}$ at a slice orientation corresponding to the rotation angle variable α. Here the z axis lies in the direction of the line $\overrightarrow{OO}_q$, and the x_1 axis is normal to both $\overrightarrow{OO}_p$ and $\overrightarrow{OO}_q$, as shown in Fig. 3.

The filter kernel $h(x_1, z_1; \alpha)$ in Eq. (17) is then given as

$$h(x_1, z_1; \alpha) = \mathcal{F}_2^{-1}\left[H\left(\omega_{x_1}, \omega_{z_1}; \alpha\right)\right]$$

and

$$H\left(\omega_{x_1}, \omega_{z_1}; \alpha\right) = |\omega_{x_2}| = \omega|\cos(\xi - \alpha)|, \tag{18}$$

where ω_{x_2} is the rotated axis from ω_{x_1} with angle α and (ω, ξ) represents the polar coordinates of $(\omega_{x_1}, \omega_{z_1})$.

The denominator of Eq. (17) is the normalizing factor that represents the sum of all the weighting coefficients of projection data sets. The projection data set $p_{\theta_p,\phi}(x_1, z_1)$ is indepenent of α, and the convolution is a linear operation. Therefore, Eq. (17) can be further simplified as follows:

$$f(\mathbf{r}) = \frac{\int_0^{\pi} d\phi \int_{-\theta_u}^{\theta_u} d\theta_p \cos\theta_p \mathcal{F}_2^{-1}\left[P_{\theta_p,\phi}\left(\omega_{x_1}, \omega_{z_1}\right) H_{\theta_p}\left(\omega_{x_1}, \omega_{z_1}\right)\right]}{4\pi^2(1 - \cos\theta_u)}, \tag{19}$$

where

$$H_\theta(\omega, \xi) = \begin{cases} 2\omega\cos\xi\sin\beta, & 0 \le \xi \le \frac{\pi}{2} - \beta \\ 2\omega(1 - \sin\xi\cos\beta), & \frac{\pi}{2} - \beta < \xi \le \frac{\pi}{2} \end{cases} \tag{20}$$

iii. Planar-integral projection reconstruction. Fourier NMR imaging techniques suggested the possibility of exciting the entire volume of an object, thereby obtaining planar-integral data sets with which an efficient volume image reconstruction could be formed.

One-dimensional planar-integral projection data along the direction T at an angle (θ, ϕ) of a 3-D object function $f(x', y', z')$ can be given as

$$p_{\theta,\phi}(z') = \iiint_{-\infty}^{\infty} f(x', y', z')\delta(T - z')\,dx'\,dy'\,dz', \quad (21)$$

where

$$T = x\sin\theta\cos\phi + y\sin\theta\sin\phi + z\cos\theta \quad (22)$$

and (x', y', z') are the rotated Cartesian coordinates. By taking the 1-D Fourier transform of $p_{\theta,\phi}(z')$ and using Eqs. (21) and (22), one obtains

$$\begin{aligned} P_{\theta,\phi}(\rho) &= \int_{-\infty}^{\infty} p_{\theta,\phi}(z')\exp(-i\rho z')\,dz' \\ &= F(\omega_x, \omega_y, \omega_z)|_{\theta,\phi}, \end{aligned} \quad (23)$$

where $F(\omega_x, \omega_y, \omega_z)$ is the image function in the spatial frequency domain and ρ is the radial spatial frequency. Equation (23) is another 3-D projection theorem, which states that the 1-D Fourier transform of the planar-integral data at angles θ and ϕ gives the 1-D radial frequency data in 3-D Fourier space.

To derive projection reconstruction using planar-integral data, let us first direct our attention to the FB method. Based on the FB method, recovery of the original volume image function $f(x, y, z)$ can be achieved through the inverse Fourier transform of Eq. (23),

$$f(x, y, z) = -\frac{1}{8\pi^2}\int_0^{2\pi} d\phi \int_0^{\pi} d\theta p''_{\theta,\phi}(T)\sin\theta, \quad (24)$$

where $p''_{\theta,\phi}(T)$ is the second derivative of $p_{\theta,\phi}(T)$ with respect to T.

2. Algebraic Reconstruction Technique and Maximum Likelihood Reconstruction

An alternative method sometimes used in image reconstruction is the iterative technique. In this category of image reconstruction, two techniques will be discussed: an algebraic method without statistical estimation and another with statistical estimation (the maximum likelihood estimation technique).

a. Algebraic reconstruction technique. The algebraic reconstruction technique (ART) was the first image reconstruction algorithm based on iterative procedures and was the algorithm first used in the EMI brain scanner developed by Hounsfield.

The ART is not the major algorithm used in commercial scanners today, however, because for practical purposes it is inefficient. It is generally agreed that the ART is slow in computation and also that iteration cannot begin before the completion of projection data collection. The use of ART based on iterative techniques, in general, is limited, with the exception of a few special cases, such as that of limited view or limited scan angle. The basic algorithm is simple and will not be discussed in detail here.

b. Maximum likelihood reconstruction. In view of the fact that many ECTs are based on statistical phenomena, a modified ART technique that incorporates the maximum likelihood estimation has been proposed. This maximum likelihood reconstruction (MLR) technique is based on the determination of an estimate $\hat{f}$ of f, maximizing the probability or likelihood $p(n^* \mid f)$ observing the measured count n^* for the unknown distribution function f.

If each emitted photon in box b, the partition of the object, is detected in detector unit d with the probability $pr(b, d)$, $d = 1, \ldots, D$, then the estimate of the unknown $f(b)$, $b = 1, \ldots, B$ using the measured counting data $n^*(d)$ is

$$\begin{aligned} \hat{f}^{(i+1)}(b) &= \hat{f}^{(i)}(b)\sum_{d=1}^{D}\frac{n^*(d)pr(b, d)}{n^{(i)}(d)} \\ b &= 1, \ldots, B \\ n^{(i)}(d) &= \sum_{b'=1}^{B}\hat{f}^{(i)}(b')pr(b', d), \end{aligned} \quad (25)$$

where $\hat{f}^{(i)}(b)$ and $\hat{f}^{(i+1)}(b)$ are the (i)th and $(i+1)$th estimates, respectively. This equation confirms that the likelihood gradually increases at each step and converges to estimate $\hat{f}(\infty)$, which has maximum likelihood. The algorithm reduces statistical noise artifacts over the Fourier transform method in ECT, such as PET, and appears to be applicable to other imaging modes.

3. Fourier Reconstruction

a. Direct Fourier reconstruction. Even though the FB algorithm is widely used because of its high image quality, it requires considerable computation, especially for the backprojection operation. One alternative algorithm is the direct Fourier domain mapping of the Fourier transform for each projection data in polar coordinates and subsequent 2-D inverse Fourier transform of the resultant Fourier domain data. This follows from the projection theorem presented in Eq. (5). Since the Fourier transform of the projection data set obtained by transaxial

scanning represents the radial set of the 2-D Fourier domain object function, it is possible to form a set of polar rasters that represent the Fourier transform of the object function through 1-D Fourier transform of the projection data. In the direct Fourier reconstruction (DFR) technique, therefore, the reconstruction procedures involve the interpolation problem as well as 2-D Fourier transform, conversion of data from polar form to Cartesian form, and 2-D inverse Fourier transform of the interpolated data with the fast Fourier transform algorithm. This algorithm would be faster than the FB algorithm, but it requires an accurate and rapid polar-to-Cartesian coordinate conversion process technique, and often the accuracy of interpolation during this conversion is of critical importance to the quality of the image.

Another important point to consider in the DFR technique is the computation time involved in interpolation; 2-D interpolation usually requires a large amount of computation. An efficient interpolation scheme, based on the concentric square raster, has been developed. This new scheme is based on the Cartesian raster suitable for the fast Fourier transform. Although the DFR technique with concentric square raster sampling seems attractive, it is not easily applicable to many practical systems, such as X-ray CT or PET, due mainly to the fixed detector spacing and sampling distance. The method is useful, however, in NMR CT, where data appear in the Fourier domain, and the sampling interval is adjustable through simple variations on the magnitude of the x and y gradient fields.

b. Direct Fourier imaging in nuclear magnetic resonance. Another interesting method in the direct Fourier transform approach applicable to NMR imaging was proposed by Kumar, Welti, and Ernst and later modified by Hutchison. This method has several distinct advantages over other methods, including the simplicity of data collection, data handling, and fast image reconstruction capability. The last is due mainly to the elimination of the time-consuming backprojection operation. Although the method itself is mathematically simple, it has to be understood in conjunction with the NMR imaging method.

In 2-D NMR imaging, by virtue of NMR, the measured signal $s(t_x, t_y)$, which is known as free induction decay (FID) or echo signal, is given by

$$s(t_x, t_y) = M_0 \iint_{-\infty}^{\infty} f(x, y) \exp(i\gamma x G_x t_x + i\gamma y G_y t_y)\, dx\, dy, \qquad (26)$$

where M_0 is the equilibrium magnetization, $f(x, y)$ is the 2-D spin density (e.g., proton density) distribution, γ is the gyromagnetic ratio, and G_x and G_y are the x and y directional gradient fields, respectively. As shown in Eq. (26), the measured signal $s(t_x, t_y)$ is the 2-D inverse Fourier transform of spin density $f(x, y)$. Therefore, the spin density image $\tilde{f}(\omega_x, \omega_y)$ can be reconstructed by taking the 2-D Fourier transform of the measured signal $s(t_x, t_y)$ as

$$\tilde{f}(\omega_x, \omega_y) = \iint_{-\infty}^{\infty} s(t_x, t_y) \exp(-it_x\omega_x - it_y\omega_y)\, dt_x\, dt_y. \qquad (27)$$

The reconstructed image $\tilde{f}(\omega_x, \omega_y)$ is then related with $f(x, y)$ as

$$\tilde{f}(\omega_x, \omega_y) = kf(\gamma G_x x, \gamma G_y y), \qquad (28)$$

where k is a constant, $\omega_x = \gamma x G_x$, and $\omega_y = \gamma y G_y$. This algorithm can also be formulated through the DFR technique discussed previously. In 3-D direct Fourier imaging, the measured signal is in the 3-D inverse Fourier transform domain. Three-dimensional images, therefore, can be obtained simply through 3-D Fourier transform operation.

These algorithm, that is, DFR or direct Fourier imaging, are based on the basic properties of Fourier transform. Direct Fourier imaging is therefore valid only for NMR imaging, because the intrinsic nature of NMR imaging lies in the Fourier transform.

B. Computerized Tomography System Configuration

1. Overview

The generic term *computerized tomography* has expanded to include several different areas, including X-ray CT, ECT, and NMR CT. CT has been applied primarily in the field of diagnostic imaging. It can be divided into three major areas, depending on the source and detector mechanism: transmission CT, emission CT, and NMR CT.

Although these systems differ slightly in terms of the energy band, the principles of physical phenomena involved, and the reconstruction algorithms employed, CT systems in general, regardless of the specific modes, comprise three basic parts: data acquisition, which includes source, sensor, and controller; processing; and display.

The data acquisition part is perhaps the most important element of a CT system, it characterizes the system, that is, defines whether it is an X-ray CT or an NMR CT, and so on. The data acquisition part includes sources, detectors, controllers, data acquisition electronics, and scanner gantry. After it receives analog signals containing image information, it converts those signals to digital form and transfers them to the computing stage.

The second part is the computer system, in which the measured and quantized data are manipulated and

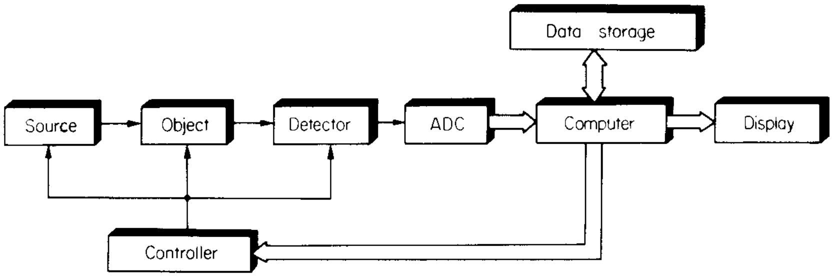


FIGURE 4 Block diagram of the general CT system.

reconstructed to create the desired images, which in turn provide information on particular structures or metabolic functions of the object under examination. The manipulation and processing of data are usually performed by general computers with array processors or by similar types of image processors, such as backprojectors, depending on the data acquisition method and reconstruction algorithm being used.

The third component is the display part, on which the observer examines and analyzes the object. A multiformat camera is commonly attached to the display console to record the pictures or images on film.

In some systems these parts are independent modular components, and in others they are integrated as a single package. A general system block diagram is shown in Fig. 4.

2. Data Acquisition

The source and sensor system with controller compose the main body of a CT system. System configuration, the related physics, and the principles of particular CTs are detailed in later sections.

It is assumed that signals containing diagnostic information can be measured in the form of electrical voltage, current, or number of photons counted. If the detected signal is not in digital form, it is converted to digital form for computer use. This process is most often performed rapidly by an analog-to-digital converter (ADC). Since all the data manipulations that follow are based on this digital representation, the performance of the ADC is vital to the overall system performance and to the final results.

An ADC is characterized by a number of parameters, including resolution, integral and differential linearities, internal noise, and operating speed. The resolution of an ADC is determined by the number of binary bits to which analog data can be digitized. For example, a 10-bit ADC has 1024 possible digitizations or output levels. Similarly, a 12-bit ADC has 4096 possible levels, while a 16-bit converter has 65,536 possible levels, and so on. The finite number of levels or steps involved in the conversion process inevitably leads to quantization errors (noise), which can be expressed as

$$\sigma_{\mathrm{q}} = \frac{\Delta}{2\sqrt{3}}, \tag{29}$$

where σ_{q} is the standard deviation of the quantization error and Δ is the step size of the ADC. It is important that the quantization noise of an ADC be kept well below the level of other sources of noise within the data acquisition system.

Integral and differential linearities are other measures of imperfection arising from the component limitations in the circuitry and circuit design of the ADC.

At present, high-accuracy and high-speed ADCs of 16-bit resolution with conversion times of less than 20 to 30 μs are widely available. Ultra-high-speed conversion with high resolution is still limited, however, by the lack of high-speed and wide-range linear ADCs.

To make full use of the dynamic range of the ADC, the input signal level should be scaled and shifted to the range of the ADC. This is usually achieved through the use of a preamplifier, attenuator, and/or level shifter. It is corrected later or during data processing through the use of appropriate software or fixed hardware devices. In some cases, special converters such as log amplifiers or integrators are used depending on the detected signal and the related processing algorithms. The output of the ADC can be stored by two methods depending on the interface technique employed: (1) It can be put into the memory of the data acquisition system until a set of data has been collected, then transmitted to the computing system. (2) It can be continuously transmitted to the computing system as the conversion operation occurs. In high-speed data transmission, the direct memory access technique is often used.

An auxiliary function of the data acquisition system is buffering, in which data are temporarily stored before being sent to the computing system. Buffering also protects the computer system from the detector electronics, which often involve high voltages.

The source and detectors are usually housed in what is referred to as the scanning gantry. The gantry is a

mechanical frame mounted in the system in a way that houses the patient during the scanning. The gantry aperture (an opening through which the patient moves in and out during the scanning procedure) varies from system to system, but is typically 45–66 cm in X-ray CT systems. In NMR CT, the aperture size is determined by the bore diameter of the main magnet, radio frequency (RF) coil sensitivity, and uniformity. A 60- to 120-cm air-bore magnet is common depending on whether the system is used for head only or whole body scans. In some CT systems the gantry can be tilted from the vertical position toward the front or back to allow for examination of specific cross sections of the patient.

3. Computer

The need for multidimensional image processing with more rapid reconstruction and higher image quality expedited the evolution of computer technology designed for imaging in general and for CT in particular. The amount of data memory and computations necessary for the reconstruction is large, and the trend is toward even larger amounts with higher resolution X-ray CT systems and with NMR CT. For example, to obtain a single slice image of 512×512 pixel size in X-ray CT, the required number of projections from different view angles is ~800 each, with more than 512 sample points. The use of a large number of sampling points is designed primarily to reduce interpolation errors. For 12-bit resolution, a memory of more than 5 megabits is needed to hold the measurement data. Reconstruction, in which convolution operations are often performed with fast Fourier transform, requires more than 40 million operations of multiplication and 250 million operations of addition. As an example, the number of required computations is summarized in Table II for both FB reconstruction and direct Fourier imaging.

The general structure of a CT computing system is illustrated in Fig. 5. In this figure, the measured data

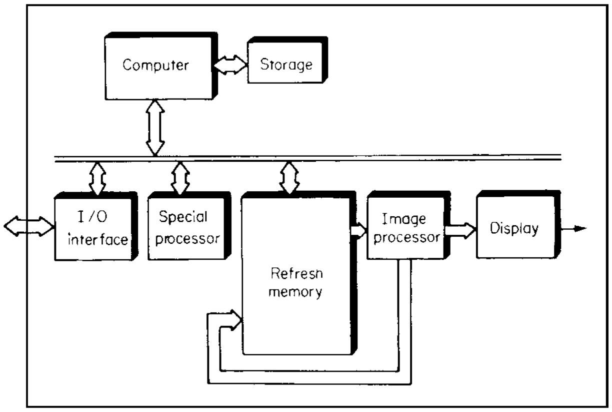


FIGURE 5 Block diagram of the general CT computational system.

are transmitted directly to a computer processing system or to an archival storage with magnetic disk, magnetic tape, or refresh memory. The measured data are transferred to a processor, either on-line or after completion of scanning, depending on the measurement speed and computer processing capability. In some CT systems, simple operations are carried out during data acquisition, and partially processed data are stored in memory or on disk. The processor can be either a general-purpose computer or special processor. The computational speed can sometimes be increased by as much as 10 to 100 times by combining special processors, such as array processors or backprojectors. They can be used most efficiently for structured data formats such as arrays or vectors.

The internal structure of the array processor comprises four functional units interconnected by internal buses (Fig. 6). The functional units are a host interface (which is system specific and provides communication with the host bus), a control processor (which controls the overall subsystem), a data memory (which acts as a data and table storage area), and a pipelined arithmetic unit (which

TABLE II Number of Computations Required for Reconstructing a Cross Section of 512 × 512 Image by Convolution Backprojection and Direct Fourier Imaging

Convolution backprojection	1024-Point FFT[a]	20,000	Multiplications and additions
	Kernel multiply	2,000	Multiplications and additions
	1024-Point IFFT[b]	20,000	Multiplications and additions
	Interpolation/view	10,000	Multiplications and additions
	Backprojection/view	262, 000	Additions
		42×10^6	Multiplications
		251×10^6	Additions
Direct Fourier imaging	Phase correction/line	2,000	Multiplications and additions
	512-Point FFT[a]	10, 000	Multiplications and additions
		11×10^6	Multiplications and additions

[a] FFT, Fast Fourier transform.
[b] IFFT, Inverse fast Fourier transform.

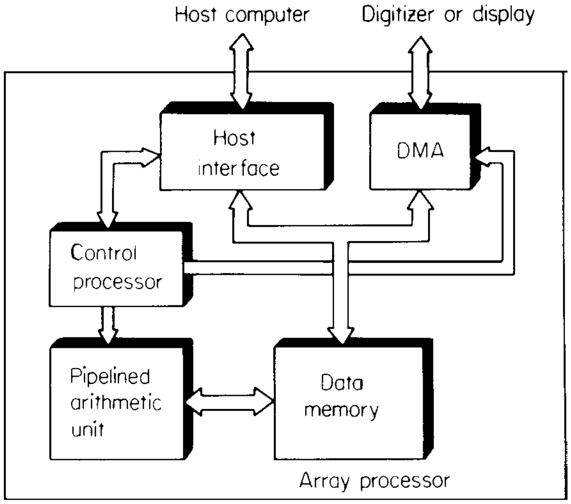


FIGURE 6 Block diagram of the internal structure of an array processor.

performs high-speed computation). In addition, an input–output interface can be used to store measured data and to transfer the reconstructed image directly to the display system without using the host computer. The speed advantage gained by the array processor is made through the parallel processing of a large number of data read from the memory of the processor, which uses its own bus, and through the use of a pipeline structure in the arithmetic unit. Distributive processing techniques can also be used to divide the processing load between the host and array processor, maximizing the efficiency of both systems.

The backprojector performs high-speed backprojection. For example, many CT systems, especially X-ray CT, are currently hindered by the backprojection operation, and therefore additional hardware computing devices like backprojectors are usually added to allow the whole image reconstruction process to be carried out almost instantaneously.

The computational speed of the special computing processor is often measured in units of million (mega) floating point operations per second (MFLOP). Through the use of array processors, more than 100 MFLOP can be easily attained. The entire processing of the data for the reconstruction depicted in Table II, for example, can be completed in only a few seconds.

Finally, reconstructed images are stored in the refresh memory, magnetic disk, or magnetic tape, depending on the amount of storage, transfer rate, access time, cost, and other factors.

Besides reconstruction, the computer provides machine control, pulse sequence control, display, and data handling. The dedicated microcomputer controller elements are therefore becoming increasingly important. They allow the implementation of highly efficient input–output operations as well as computational structures. The host computer therefore becomes an interface with the operator and data acquisition elements in the CT scanner and related data bank.

4. Display

The display of the CT system provides a means of visual examination of the scan information obtained through the data acquisition system and computer. The CT display system is the fundamental link between the CT image and the human visual system, which provides the radiologist with maximal diagnostic information. Display monitors are usually attached to an operator console through which all the commands are made.

To facilitate visualization of small differences in tissue density, the value from each address of the display memory is first modified by some transforms before being changed to an analog signal by a digital-to-analog converter. Most present-day displays utilize 256 gray levels in accordance with human visual perception characteristics. In conventional CT display systems, a selected range of CT values is displayed uniformly over the range, which is preselected by setting window level and width. The window level control determines the midpoint of the density range, while the window width determines the width of the range. By changing the window width and level, the observer can enhance the visualization of the subject structures. In essence, the window width controls the contrast of the image, and the window level changes the density of the tissue to be displayed. Decreasing the window width usually results in a high-contrast image, but it also enhances noise.

In display systems, image processors are usually utilized to facilitate the manipulation of image enhancement and restoration (including filtering, selection of regions of interest, cut-view display, zooming, and distance measurement) without interrupting the host computer.

III. AREAS OF APPLICATION

A. X-Ray Computerized Tomography

X-Ray CT is a product of X-ray technology with advanced computer signal processing and capable of generating a cross-sectional display of the body. It is, in fact, the origin of the entire CT evolution, which began in the early 1970s.

The basic principle of X-ray CT involves X-ray generation, detection, digitization, processing, and computer image reconstruction. As the X rays pass through the body, they are attenuated at different rates by different tissues. The attenuated X rays are then collected by detectors and converted to digital impulses or data by the ADCs. These

digital data are fed into a computing device for image reconstruction.

For simplicity, the pencil beam type of X ray is scanned along a line at a given direction or view. To achieve several different angles or perspectives, this scanning is repeated at each given angular view simply by rotation of both the X-ray tube and detectors.

1. Basic Physical Principles

a. Contrast mechanism and projection data. The photon density that emerges when a narrow beam of monoenergetic photons with energy E and intensity I_0 passes through a homogeneous absorber of thickness x can be expressed as

$$I = I_0 \exp[-\mu(\rho, Z, E)x], \tag{30}$$

where μ, ρ, and Z are the linear attenuation coefficient, density of the absorber, and atomic composition or number, respectively.

In the energy region where most commercial X-ray CT systems are being engaged for medical tomography (~70 keV), two types of interactions are dominant: photoelectric absorption and coherent and incoherent scattering.

In photoelectric absorption, the X-ray photon is completely absorbed by transferring all of its energy to an electron. Scattering is subdivided into two components: Rayleigh (coherent) scattering and Compton (incoherent) scattering. In Rayleigh scattering, the direction of the X-ray photon is changed but the energy is conserved. In Compton scattering, on the other hand, both direction and energy are changed. Figure 7 depicts the interactions of the X rays with water. It shows the contribution of each photon interaction to linear attenuation coefficients of different types as a function of energy. If the absorber is not homogeneous, $\mu(\rho, Z)$ is simply a space-variant function dependent on the material. By directing a monochromatic X-ray beam in the y direction, the output X-ray intensity $I(x)$ is

$$I(x) = I_0(x) \exp\left[-\int \mu(x, y)\, dy\right], \tag{31}$$

where $I_0(x)$ and $\mu(x, y)$ are the incident X-ray intensity and X-ray attenuation coefficient, respectively. By taking the logarithm and rearranging Eq. (31), one can obtain projection data $p(x)$ as

$$p(x) = -\ln[I(x)/I_0(x)], \tag{32}$$

where $p(x)$ is a simple integration or summation of the total attenuation coefficients along the X-ray path. In X-ray CT, the contrast is the difference in attenuation coefficients of the material involved. Since each set of projection data

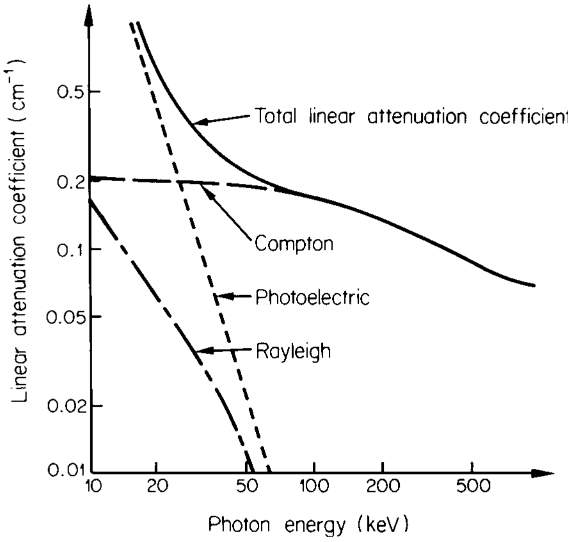


FIGURE 7 Linear attenuation coefficient of water and contribution of each interaction to the attenuation of X rays as a function of the energy.

represents the integral value of the attenuation coefficients along the path, the projection data taken at different views are the basis for tomographic image reconstruction.

b. Beam hardening. X-Ray beams generally used in X-ray CT are not monoenergetic and have a finite spectrum. When this polychromatic X-ray beam passes through a material, X rays of different energy in the spectrum undergo different attenuation, and as a result the output energy spectrum differs from the input spectrum; that is, lower energy X rays are attenuated more heavily than higher energy X rays. This trend is accentuated if the path length is large or if material possesses components of high atomic number. The consequence of this nonuniform attentuation of the polychromatic X-ray beam is called the *beam-hardening effect*. It produces a visible artifact in the final reconstructed image. The original EMI scanner utilized compensatory measures to deal with the beam-hardening effect. A water bag was used as a compensator by surrounding the head so that the total path lengths of the X rays were always the same for all projections. Another method is to preharden the X-ray beam by passing it through an aluminum or copper filter so that the output X ray is close to the monoenergetic beam before it passes through the body.

c. X-Ray source. Two types of X-ray sources are currently used in X-ray CT. The most simple type is the fixed-anode X-ray tube, in which the anode is cooled by oil and is continuously energized. A typical focal spot size of this type is about 2 mm × 16 mm on a 20° angle tungsten target. The relatively small heat dissipation capability

associated with this type of tube limits the amount of photon flux generation, resulting in a statistically noisier image than others, for example, the rotating-anode X-ray tube, when operated continuously. The rotating-anode X-ray tube, on the other hand, allows for a large photon flux because it has much greater heat capacity.

d. Detectors. X-Ray photons are collected by radiation detectors of various kinds, such as a scintillation crystal coupled with a photomultiplier tube (PMT). In general, the output of the detectors consists of electrical signals that are proportional to the incident X-ray energy or fluence. The most important parameters to be considered in the selection of detectors for X-ray CT are efficiency, response time (or afterglow), and linearity. Efficiency refers to the absorption and conversion efficiency of the incident X rays to electrical signals. Linearity refers to the dynamic range of the detector response. Response time refers to the speed with which the detectors can detect X-ray photons and recover to detect the next photon. This is determined by the afterglow—one of the important characteristics of detector materials for X-ray CT application.

Detector types currently in use for X-ray CT can be divided into two classes: scintillation detectors and gas ionization detectors.

i. Scintillation detector. Scintillation crystals [such as NaI(T1) and CsI(T1)] produce flashes of light as they absorb X-ray photons. The light is then converted to electrical signals by subsequent electronics. The following two types of scintillation detector systems are commonly used in X-ray CT:

Scintillation Crystal–PMT Coupled Detector. Light produced in the crystal is coupled to the photocathode of a PMT. In the PMT, photoelectrons are generated from the photocathode as the light strikes it. These electrons are multiplied through a series of cascaded dynodes in which electron multiplication processes take place. Each dynode produces more electrons than incident electrons. The multiplied or amplified electrons constitute output signals in the form of a charge or a current, which are the indicators of the energy and fluence of the incident X-ray photons on the scintillation crystal. An overall gain of a few millions is common in most PMTs. Other detector crystals, such as BGO (bismuth germanate) and CaF_2, are also used.

Scintillation Crystal–Photodiode Coupled Detector. The performance of this detector, which was developed more recently than its counterparts, has been found to be satisfactory. A typical scintillation crystal–photodiode detector comprises a CsI(T1) scintillation crystal, a P–N junction photodiode coupled to the crystal, and a preamplifier for the low-level signal amplification.

The incident X-ray photon is converted to visible light in the scintillation crystal, which then falls into the PN junction photodiode. Generated electron-hole pairs are collected at the junctions. Since the generated current is usually weak, a low-noise preamplifier is required. The voltage output is proportional to the energy and fluence of the X-ray incidence on the detector.

ii. Gas ionization detector. Some commercial X-ray CT systems use gas ionization detectors. To improve the detector efficiency, xenon gas is pressurized by as much as 20 atm and filled in a long chamber to maximize detection efficiency. It consists of tungsten plates, which serve as anodes for electrons. When X-ray photons are incident on the detector cell, gas is ionized. This ionized gas provides current that is directly proportional to the X-ray photon energy. The advantages of the gas ionization detector are high spatial resolution capability and simplicity. In addition, the compact detector assembly can be made on a large scale by packing a few hundred equivalent detector elements that have detector widths as narrow as 1 mm. The disadvantage of the ionization detector, even with highly pressurized gas, is the low detection efficiency. Characteristics of each detector are briefly described in Table III.

e. Data acquisition and reconstruction algorithms. At each view, detector signals from the detector system are converted to digital pulses by the ADC. These signals are collected by the computer via signal processing

TABLE III Characteristics of Detectors Used in X-ray CT

Type	Advantages	Disadvantages
NaI(T1)-PMT	High detection efficiency, 100% at 70 keV(1-in. crystal)	Afterglow; restricted dynamic range; low packing density; hygroscopic
CaF_2- PMT	No afterglow	Low detection efficiency (62% with 1-in. thick)
BGO (bismuth germanate)-PMT	No afterglow; high detection efficiency; nonhygroscopic	Low light output
CsI(T1)-photodiode	Good spectral match with available PMTs; high detection efficiency (94.5% at 120 keV with 5-mm thick)	—
Xenon gas ionization detector	Simple and compact; no afterglow; high resolution capability	Low efficiency; possible instability; slow response time

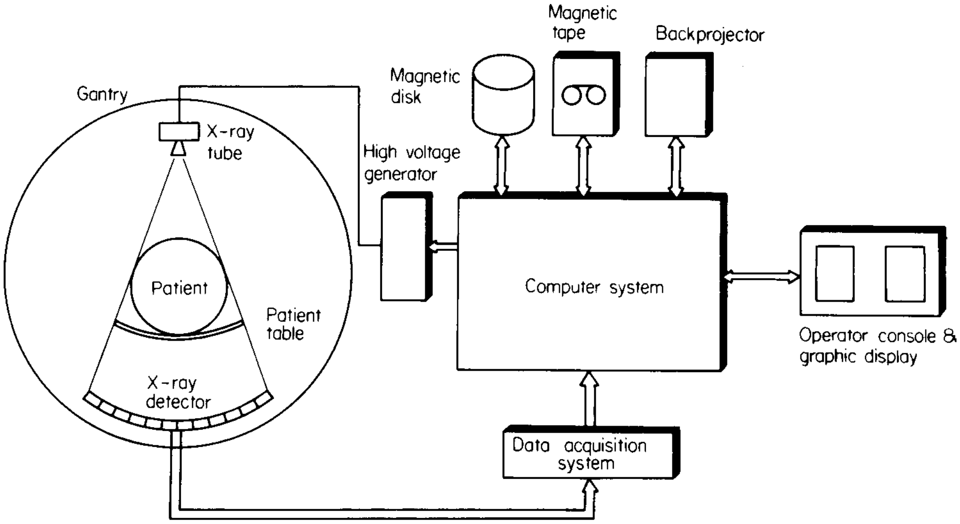


FIGURE 8 X-Ray CT system configuration.

electronics for image reconstruction. The steps of the view angle between successive views are normally ~1° or less, and a few hundred views are usually taken from each tomographic slice. After the projection data sets (i.e., line-integral projection data) are obtained, image reconstruction takes place. The time required for image reconstruction can be relatively long in comparison to data acquisition time. Therefore, the procedure usually requires special-purpose computer peripherals, such as array processors or backprojectors.

f. System configuration. Figure 8 is a block diagram of a typical X-ray CT system currently in use. A 16- to 32-bit minicomputer system equipped with a dedicated backprojector is normally used for data acquisition, signal processing, and system control.

2. System Evolution

Although remarkable progress has been made in all aspects of X-ray CT since it was introduced in 1972 by Hounsfield at EMI, the most significant changes have been made in the area of data acquisition. These stages of progress are classified into "generations." Since 1972, the X-ray CT has evolved from the first generation to the fourth and possibly to the fifth generation, which will incorporate the latest developments in dynamic scanners.

a. First generation. The first generation naturally entailed the first EMI scanner developed by Hounsfield. This scanner used a single pencil beam and a single detector, which translated and rotated synchronously. There was translational motion across the object being scanned, and at the end of each translational motion an incremental 1° rotation followed in preparation for the upcoming scanning. This procedure is depicted in Fig. 9a. The collection of projection data needed for tomographic image reconstruction of a slice took several minutes.

b. Second generation. The design of the second-generation CT incorporated a narrow-angle fan beam X ray and an array of multiple detectors. Since the diverging fan beams passing through the patient increase data collection channels, the number of angular rotations required could be reduced. Therefore, the scan time in this second generation was shortened substantially; the nominal scan time was ~20 s. The second-generation scanner still entailed translational motion as well as rotational motion to cover the object fully, however. Figure 9b illustrates the configuration of the second-generation CT scanner.

c. Third generation. In the third generation, the fan beam angle is widened, and the fan covers the entire object to be scanned. Each projection path is defined by a matching detector, which can be either a small and narrow scintillation detector slab or a segment of a gas ionization chamber. Because the entire object is covered or encompassed, no translational movement is required. Therefore, both the X-ray tube and the detector array need only simple rotational motion around a fixed axial center. The entire 360° is usually scanned for whole data collection. Scan time is as short as 3 s. This high speed scan capability allows for synchronized rotational motion with physiological signals such as that obtainable from electrocardiograph (ECG), when the imaging of moving organs is required. The major drawback of this configuration is that the effects of the drift of the detector are cumulative so that artifacts

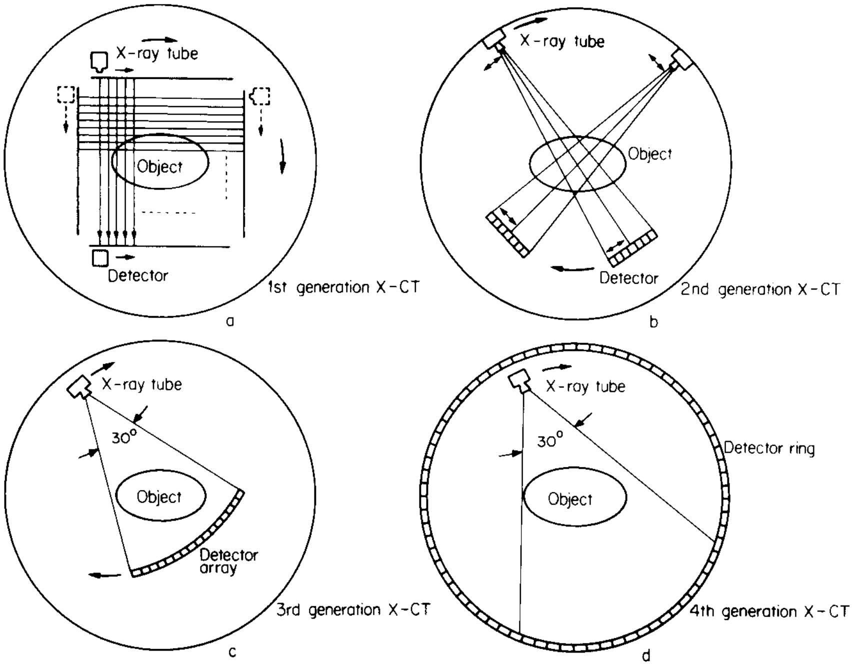


FIGURE 9 X-Ray CT system evolution from first- to fourth-generation scanners. (a) First generation; (b) second generation; (c) third generation; (d) fourth generation.

appear on the reconstructed image. Figure 9c shows the configuration of the third-generation CT scanner. Almost all third-generation scanners use pulsed X-ray sources to take advantage of significant dead time between successive views.

d. Fourth generation. The construction of a stationary circular ring detector array is probably the ideal choice of the detector for the rotating X-ray source. A striking analogy is the circular ring PET scanners of various types developed during the 1970s. This stationary ring is the most distinct feature of the fourth-generation X-ray CT. The X-ray source rotates, but the detector array does not. A wide-angle fan beam X ray encompasses the entire patient, and 600 or more stationary detectors form a circular ring array. With this kind of configuration, detector drift is not cumulative and therefore can be corrected. The advantages of this system are similar to those of the third-generation systems, but the main drawback of the third-generation systems—drift effects—has been eliminated since detector drifts no longer accumulate over successive views. The fourth-generation systems are generally more expensive, however, due to the large number of scintillation detectors and PMTs employed. Figure 9d shows a schematic diagram of the fourth-generation X-ray CT scanner.

e. Dynamic scanner. For the imaging of moving organs, such as the heart, ultrafast scanners are required. Typical systems of this kind are the dynamic spatial reconstructor (DSR) and cardiovascular CT (CV CT). These can be categorized as the fifth generation X-ray CT scanners.

i. Dynamic spatial reconstructor. The development of the DSR at the Mayo Clinic was completed in 1982. It can produce real-time images of body organs in motion. The DSR comprises 28 X-ray sources with 28 opposing X-ray imaging chains, which are image intensifiers coupled with X-ray detection phosphors mounted in the same gantry. The physical size of components and the required radiation flux determine the number of X-ray sources and imaging chains. As many as 240 images of adjacent slices with a thickness of 1 mm can be made from a cylindrical volume that is 38 cm in diameter and 24-cm long. A complete volume scan is achieved in 0.01 s after each of the 28 X-ray sources is pulsed in succession (for 0.34 ms). These scans can be repeated 60 times per second. A high temporal resolution image can be obtained with 28 angles of view recorded in 0.01 s if the gantry is kept stationary. For stationary objects, however, high spatial and density resolution images can be produced using all 240 views for reconstruction. Trade-offs between temporal, spatial, and density resolution can be made by selecting the appropriate subsets from the total projection data. Figure 10 depicts a schematic diagram of the DSR scanner.

ii. Cardiovascular computerized tomography. This system was proposed and developed by Boyd *et al.* at the University of California, San Francisco, with the same goals in mind, but with a more compact and physically

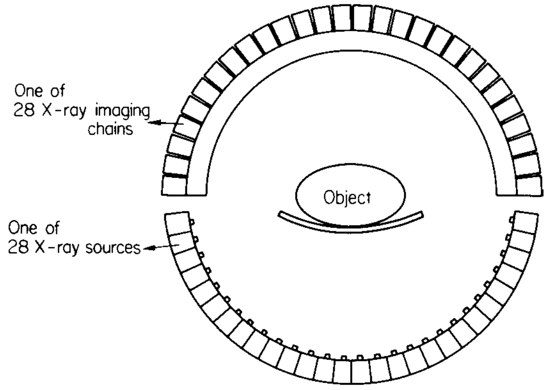


FIGURE 10 First dynamic scanner DSR developed by the Mayo Clinic.

integrated design than that of the first DSR. The heart of CV CT is the electron beam scan tube and stationary scintillation crystal–photodiode coupled detector array. An accelerated and focused electron beam is deflected by a computer-controlled bending magnet to be swept along a 210° curved tungsten target ring. Four target rings are swept serially to obtain a multiple section examination. Approximately 30° of the fan-shaped sector of the X-ray beam generated at the tungsten target is detected by the detector array for image reconstruction. The detector array comprises two detector layers of a half-ring. A simple scan produces two side-by-side tomographic slices; a total of eight slices can be produced to cover a region ~9-cm deep by sweeping four targets in succession. Two adjacent tomographic slices can be obtained in 50 ms and eight slices can be obtained in 200 ms if four targets are swept serially.

Although the basic principles of X-ray CT have not changed since they were introduced, the scanning scheme, speed, and performance of X-ray CT have improved substantially. At the same time each stage of X-ray CT development has brought about new applications and widened the scope of X-ray CT. X-Ray CTs of different forms are now utilized in many diverse fields, including the inspection of tires on the production line and the detection of defects in nuclear reactor cores.

Scanning speed has evolved from several minutes to a few milliseconds, making possible dynamic scans of moving organs, such as the heart. The search for better and higher contrast and spatial resolution with shorter and shorter imaging time is expected to continue, and the growth of applications of X-ray CT to fields other than medical imaging is anticipated.

B. Emission Computerized Tomography

Emission computerized tomography is an imaging technique capable of visualizing the 3-D distribution of radionuclides in the human body. It is also capable of measuring quantitatively *in vivo* biochemical and metabolic functions. This is in contrast to transmission X-ray CT, which emphasizes visualization of the anatomical structure of the human body; that is, spatial resolution and tissue density contrast are more important. The field of ECT is one of the fastest growing areas in nuclear medicine and is rapidly replacing the conventional 2-D projection images known as scintigraphy.

Unlike transmission X-ray CT, ECT generally suffers from poor resolution or statistics. The number of photons that can be detected is usually limited, especially for the tomographic mode operation, in which the quantity of photons available is limited to a small columnlike volume for each detector or detector pair. In addition to a lack of sufficient photon statistics, emission CT is hampered by the fact that in the detection process photons emitted by the body undergo attenuation in the tissue. To provide a reliable reconstructed image, attenuation due to body tissue must be corrected.

ECT can be conveniently divided into two modes—positron emission tomography (PET) and single photon emission computerized tomography (SPECT)—according to whether the radionuclides employed undergo decay with the emission of positrons, which provide annihilation photons, or with the emission of γ rays. The former is generally considered to be more suitable for the tomographic mode, while the latter has been more readily available and has begun to show potential in tomographic imaging. In this section, these two modes, as potential tomographic imaging modalities, are detailed from a physical point of view.

1. Positron Emission Tomography

Although the potential of positron imaging was recognized as early as the 1950s, its actual tomographic mode imaging began only after X-ray CT was developed in 1972. PET is a brilliant example of a joint effort of many disciplines, including physics, electronic instrumentation, and computers. PET requires short-lived cyclotron-produced radionuclides and the appropriate labeling of these radionuclides to suitable radiopharmaceuticals that can be administered to the human body, as well as instrumentation for the detection of annihilation photons and appropriate signal processing to provide high-resolution images. Small cyclotrons are now available for use in medical facilities. Advances in the rapid chemical synthesis of radiopharmaceuticals now permit a large number of labeled compounds to be used in positron imaging. Above all, there has been phenomenal growth in PET instrumentation research and development in recent years; a few system designs are potentially capable of

imaging with a resolution as high as 2 to 3 mm full width at half-maximum (FWHM).

This section is devoted to the instrumentation aspects of PET and is divided into three parts. The first part introduces some basic principles, the second deals with various physical factors affecting system performance, and the third discusses the evolution of PET system designs.

a. Basic principles.

i. Positron emitters and physics. Positron-emitting radionuclides possess several important physical properties that make PET a unique imaging technique. Its most important property is the directionality and simultaneity of photons generated by annihilation. The emitted positron combines with a nearby electron, and two photons are generated by a phenomenon known as annihilation. These two annihilation photons, each with energy of ~511 keV, are generated simultaneously (simultaneity), and they travel at ~180° from one another (directionality). The nearly collinear direction of the two annihilation photons makes possible the identification of the annihilation event or the existence of positron emitters through the detection of two photons. This is usually achieved by a coincidence detection circuit that records an event only if both detectors sense annihilation photons simultaneously (solid lines, Fig. 11). Because the two detectors record coincidence events only from a volume of space defined by a column or strip joining the two detectors, the physical collimation used in SPECT to confine the direction of incoming photons can be eliminated. For this reason, coincidence detection is also called electronic collimation. The total number of coincidence events detected by a given pair of detectors constitutes a measure of the integrated radioactivity (i.e., line-integral projection data) along the strip joining two detectors. From a complete set of line-integral projection data obtained from several views with detectors surrounding an object, the activity distribution within the slice can be reconstructed by using the algorithms discussed in Section II.A.

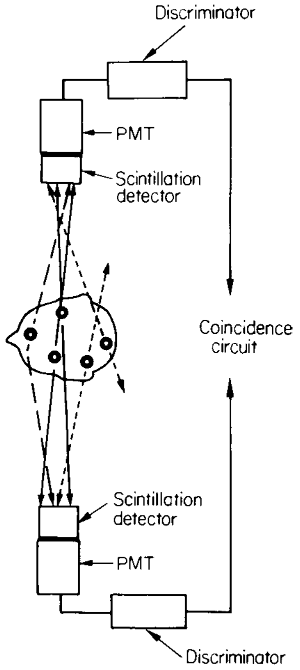


FIGURE 11 Principle of coincidence detection. True coincidence (solid line), random coincidence (dashed line), and scattered coincidence (broken line) are indicated.

The accuracy of the spatial localization of a positron-emitting radionuclide by the coincidence detection procedure is limited, however, by two physical properties of positron annihilation: (1) The two annihilation photons are not exactly collinear, thereby creating angular uncertainty, and (2) there is uncertainty about the annihilation position of the emitted positrons, that is, uncertainty about positron ranges. The latter is strongly dependent on the kinetic energy of the emitted positron. The combined effect of these two factors introduces a fundamental uncertainty in locating the sources of the positrons. This uncertainty, depending on what kind of radionuclides are used and on detector separation, is typically about 2 to 3 mm FWHM. This value is accepted as a lower limit on the resolution that can be achieved by a positron camera. Despite these inherent spatial resolution limits, PET could potentially be the highest resolution nuclear imaging technique.

ii. Advantages of PET. Perhaps the greatest advantage of PET over other nuclear imaging systems is the electronic collimation with which a solid angle is extended to virtually the entire object. The sensitivity (photon collection capability) gain due to this large detection solid angle that results from electronic collimation is greater than that of SPECT. This is considered to be an important advantage of PET over SPECT.

Other advantages of PET, which uses electronic collimation, are the uniformity of resolution and sensitivity over the entire range. The latter, uniform sensitivity, stems from the fact that the combined attenuation affecting a photon pair is the same, regardless of the position of the annihilation, as long as the annihilation points are inside the column defined by a detector pair. This allows us to compensate accurately for the attenuation suffered within the object. This is again a distinct advantage of PET over SPECT.

Another advantage of PET consists of the physiological aspects of most of the available positron-emitting radionuclides, which are usually of low atomic number. Among these radionuclides, ^{11}C, ^{13}N, ^{15}O, and ^{18}F are used most often in PET because of their physiological affinity in the human body and their short physical half-lives

(^{11}C, 20.34 min, ^{13}N, 9.96 min; ^{15}O, 2.05 min; ^{18}F, 110 min), which facilitate effective imaging with minimal dosages to the patient. The radionuclides ^{11}C, ^{13}N, and ^{15}O are the major components of molecules in living matter. Therefore, they are closely related to the metabolic processes in human physiology.

b. Physical factors affecting system performance.

i. Detector and related materials. The spatial resolution of the positron camera depends on the width of the detection channel and the sampling interval. The former sets the resolution limit obtainable with the system, provided that the sampling requirement or the Nyquist sampling criterion is met. The detection channel can be characterized by the detector aperture function and can be determined by simple ray tracing. At the center of a detector pair, it can be represented by a triangle with an FWHM equal to half the detector width. The aperture function toward the detector becomes trapezoidal and ultimately rectangular at the position immediately adjacent to the detector.

In most PET instruments, a standard γ-ray detector design is chosen for detector assembly. It consists of a PMT and a scintillation crystal coupled together.

Several scintillation crystals used in the past are being recommended for PET design. In the early 1970s, NaI(T1) was the most commonly used scintillation detector for PET, even though it was comparatively inefficient in high-energy 511 keV photon detection. A more recent development in scintillation detectors was the introduction of BGO, which led to substantial improvement in detection efficiency. The nonhygroscopic nature of BGO crystals further facilitates compact detector assembly or packing, which also leads to an overall increase in detection efficiency. The design of a high-resolution PET system was made possible by the introduction of BGO, since an extremely narrow slab of each detector crystal (4–5 mm in width) is considered to be the essential requirement of the high-resolution PET.

Two other crystals have been considered for PET application, CsF and BaF_2, both of which have very rapid detection capabilities. Therefore, these crystals have made excellent time resolution feasible. This advantage has been exploited in time-of-flight (TOF) PET systems.

Some of the physical properties of the detector materials mentioned here are listed in Table IV.

ii. Sampling. Because detector width is limited by detection efficiency, the widths of the detectors are usually kept as large as the resolution allows. With a given detector size, maximum resolution can be achieved by satisfying the Nyquist sampling criterion; that is, the sampling distance must be less than half the distance of the highest spatial resolution obtainable by that particular ring PET design, (i.e., half the detector width). In projection data obtained with a detector width of w, a maximum spatial frequency is about $2/w$, because the intrinsic resolution expected is about half the detector width. The desired sampling distance, therefore, should be less than $w/4$.

In the hexagonal- or octagonal-geometry PET system, desired sampling can easily be achieved by introducing translational and/or rotational motions (Fig. 12a). In the circular ring system, however, improving the sampling arbitrarily has been difficult, and several sampling schemes for overcoming this inherent difficulty have been proposed and implemented in experimental systems. For example, wobbling motion has been widely used (Fig. 12b). However, two sampling schemes suitable to the circular ring

TABLE IV Properties of Detector Materials Used in Positron Cameras

Detector material	NaI(T1)	BGO	CsF	BaF_2
Density (g/cm^3)	3.67	7.13	4.64	4.89
Linear attenuation coefficient at 511 keV (cm^{-1})	0.34	0.92	0.44	0.47
Scintillation decay time (ns)	250	300	5	0.8
Emission wavelength (Å)	4100	4800	3900	2250
Energy resolution in FWHM at 511 keV (%)	>7	>10	23	13

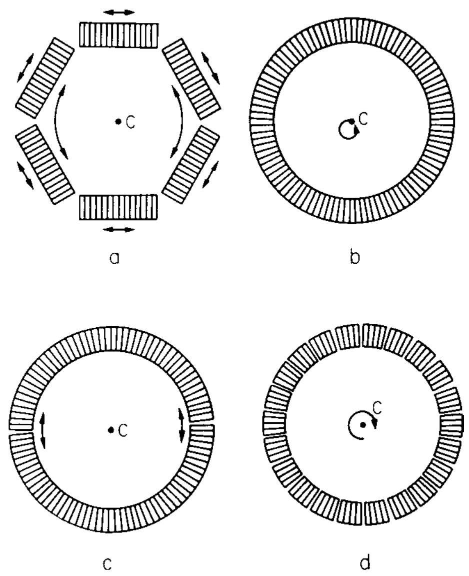


FIGURE 12 Sampling motions. (a) Hexagonal geometry with translational and rotational motions. (b) Wobbling motion applied to circular ring systems. In this scheme, the entire ring moves in a circular path (small track of circle shown at the ring center); this is 2-D motion. (c) Dichotomic motion applied to circular ring systems. In this scheme, two half-arcs, which form a ring, move back and forth in opposite directions along the circular path. Note that the motion is 1-D. (d) Positology. In this case, continuous rotation of a ring with unevenly spaced detectors provides finer sampling.

system have been developed, namely, dichotomic sampling and clam shell sampling.

Although the wobbling scheme is one of the most common methods, it is generally believed that the samplings obtained by this method are usually neither uniform nor equally spaced. To obtain more uniform and equally spaced samplings with the fewest possible number of motions, a new sampling scheme, known as dichotomic sampling, has been proposed and incorporated into an experimental system (Fig. 12c). This scheme employs two half-rings (from which the term *dichotomic* is derived), which rotate in such a way that finely sampled parallel or fan data sets can be obtained with a minimal number of scan stops. Another scheme, developed at the University of California, Berkeley, is similar to the dichotomic sampling scheme but employs a slightly different motion.

Another method for improving the sampling scheme is the positology (Fig. 12d) developed by Tanaka *et al.* in Japan. In this scheme, rapid rotational motion is employed with a nonuniformly spaced circular detector array. With this scheme, through the rotation of the entire ring, the desired sampling can be achieved.

iii. Sensitivity. Sensitivity in PET is defined as the capability of detecting the true coincidences (solid line, Fig. 11) with a given amount of radioactivity. Sensitivity has traditionally been measured with a phantom of diameter $d = 20$ cm, filled with a uniform activity concentration ρ (μCi/cm^3) (1μCi = 37,000 disintegrations per second). By considering several factors, including the activity in the field of view, the self-attenuation of γ rays within the phantom, the solid angle subtended by the detector array, and the detection efficiency of the array, an empirical formula for sensitivity measure is derived and given by

$$C_t = 14{,}500\rho\alpha\varepsilon^2 h^2 d^2/D \quad (\text{counts s}^{-1}\ \mu\text{Ci}^{-1}\ \text{cm}^{-3}). \tag{33}$$

In this equation, α is the probability of no scatter of both annihilation photons, ε is the detector efficiency including the detector packing ratio, h is the thickness of slice to be imaged, d is the diameter of the phantom, and D is the ring diameter.

If the system contains two or more detector rings, coincidence can be measured between the detectors in the different rings (coincidences not in the planes that are perpendicular to the axis) to increase the sensitivity of a given imaging plane. Sensitivity is higher in the cross-slice planes because they involve twice as many detectors.

iv. Random coincidences. As the source of background noise in PET images, random or accidental coincidences occur when two photons emitted from two independent positions are detected within the coincidence resolving time τ (dashed line, Fig. 11). Random coincidences produce a haze of background over the field of view in the reconstructed image.

For a uniform distribution of activity, a formula for the random count rate has been established and given by

$$C_r = \tau f_d C_s^2 \qquad (\text{counts s}^{-1}\ \mu\text{Ci}^{-1}\ \text{cm}^{-3}), \tag{34}$$

where C_s is the single count rate for the entire ring, τ is the coincidence resolving time, and f_d is the fraction of detectors covering the object in the whole detector ring in coincidence with any one given detector, that is,

$$f_d = (2/\pi)\sin^{-1}(d/D). \tag{35}$$

Because both the single rate and true coincidence rate are proportional to the amount of activity, it is apparent from Eq. (34) that the random coincidence rate is proportional to the square of the true coincidence rate.

To reduce the random coincidences, it is imperative to minimize single counts. Because single counts can arise from both in-slice and out-of-slice annihilation events, they can be suppressed by limiting the detection channels by slice collimation or by the increase in the energy threshold so that maximum rejection of any scattered single events can be achieved. It is common practice to reduce as many annihilation events as possible from out of slice by using annular interdetector ring collimators.

Random coincidences can be partially corrected by software, provided that the single counts of each channel and the coincidence resolving time are recorded.

v. Scattered coincidences. Scattered coincidences occur when one or both γ rays resulting from an annihilation event are scattered in the medium and detected with the remaining energy of the γ rays above the energy threshold (broken line, Fig. 11). They give incorrect positional information and produce a line-spread function with long tails. The number of scattered coincidences can be reduced by setting the energy threshold level high and by using tighter interslice collimation.

vi. Correction of random and scattered coincidences. Random coincidences can be corrected by two simple procedures. One is by using the delayed coincidence measurement with the same time window and the other, mentioned earlier, is by estimating random coincidences using Eq. (34). As stated above, random coincidences can be reduced either by minimizing the coincidence time window or by reducing the activity.

Since the scattered coincidences are an inherent physical property, they cannot be easily reduced or eliminated as random coincidences. Because it is prompt in nature, differentiation between true and scatter simply by minimization of the time window is therefore difficult. In addition, the energy loss in small-angle scatter is so small that it is difficult to differentiate through the energy window. In the case of multilayer ring geometry in imaging multiple

slices, scatter elimination is achieved by the use of tighter interslice collimation. However, one should admit some loss of observable volume when large-volume imaging is of importance.

vii. Attenuation correction. Attenuation correction is one of the most important parts of PET imaging. A variety of correction schemes have been developed in the past. One of the simplest and crudest ways of correcting attenuation is through the use of geometric shape, that is, by finding edge contours and using contour information to determine attenuation lengths for the subsequent correction. In this case, attenuation coefficients are customarily assumed constant. A more accurate method is to use the transmission scan information obtained by an external positron source surrounding the patient. This technique often suffers from statistical noise and thereby makes attenuation correction difficult, unless a sufficient amount of data is taken. The correction can also be made through X-ray CT by a procedure similar to that described earlier, but one should be aware that attenuation coefficients are different from those obtained with 511-keV photons.

c. Evolution of positron camera development. There have been continuous efforts to develop PET imaging systems since the early 1950s. Some notable examples include two NaI detector systems developed by Brownell and Sweet in the early 1950s, 32 discrete NaI(Tl) detector systems developed by Rankowitz *et al.*in 1962, and the PC-I developed by the Massachusetts General Hospital (MGH) with two banks of detectors (127 detectors per bank) in 1972. Although these systems ultimately were intended for tomographic imaging, they remained quasi-tomographic machines until the introduction of the X-ray CT scanner by Hounsfield in 1972.

The first tomographic systems developed after the introduction of Hounsfield's model were the PETT (positron emission transaxial tomograph) I–III series designed by Ter-Pogossian and Phelps of St. Louis and CRTAPC (circular ring transaxial positron camera) designed by Cho *et al.* at UCLA. As a hexagonal detector array, the former employed both translational and rotational motions, while the latter remained virtually stationary. Since these developments, PET instruments have rapidly improved through the formulation of a variety of new concepts. For example, system geometry has evolved from planar to hexagonal type, from hexagonal to circular type, from single ring to multiring, and so on, Detector material has been changed from NaI(Tl) to BGO, CsF, or BaF_2. The time of flight (TOF) technique has also been introduced. Also, several commercial companies have begun to design PET systems.

i. System geometry. It is interesting to observe how system geometry has evolved since the mid-1970s. System geometry (the arrangement of detector arrays) is the most basic design choice, because it determines fundamental system performance. It can be categorized into three basic types: planar, polygonal, and circular ring. To cover the imaging volume in the axial direction, multiring systems have appeared in which several rings are stacked on one another. They provide high total sensitivity and offer $2N - 1$ (N is the number of rings) image slices simultaneously.

The trend appears to be toward more generalized circular ring types of various forms. Among the advantages of circular ring geometry are uniformity, high sensitivity due to high packing fraction, and high angular sampling capability. Because the requirement for angular sampling can be met even by the stationary ring, systems do not require further rotation to improve angular sampling. Although linear sampling has been a limiting factor in the circular ring system, various solutions to this problem have also been suggested, and their efficacy has been proved experimentally. It is also worth noting that trends are in the direction of volume imaging, either through multilayer rings or through a spherical-PET (S-PET) system to be described later. Yet another avenue, one that requires further technological development, is the TOF approach. This is considered to be an adjunct to improve system resolution.

ii. Planar system. The planar type consists of two detector planes (made of either discrete crystal arrays or position-sensitive devices, such as the Auger camera or multiwire proportional chamber) facing one another. A set of projection data can be obtained by rotating the dual planes around the patient, from which a series of image slices covering sufficient axial volume is reconstructed. A disadvantage of this type is the large number of angular rotations it requires.

MGH systems (PC-I, PC-II) and the commercial version of those systems (TCC 4200) fall into the category of discrete crystal arrays.

iii. Polygonal system. Such systems as PETT III and PETT IV by Ter-Pogossian and Phelps and ECAT and NeuroECAT by Ortec fall into this category. With the exception of the octagonal-shaped NeuroECAT, these systems are hexagonal. In this type of system, coincidence detection channels are formed between those banks opposing one another so that ring efficiency is usually limited, particularly toward the periphery of the image. An advantageous feature of the hexagonal system is the relatively simple translational and rotational sampling motions, which fulfill the requirements of uniform linear and angular samplings.

iv. Circular ring systems. A natural extension of the polygonal PET system is circular ring geometry, which provides uniformity as well as natural symmetry. The first circular ring PET system was conceived and developed by Cho *et al.* at UCLA in 1975. Various other circular

ring systems were developed subsequently by Budinger *et al.* at Berkeley, by Bohm *et al.* at Stockholm, by Carroll *et al.* at Cyclotron Corporation (also at Berkeley), and by various commercial firms. Some of the undersampling problems associated with ring systems have been resolved by incorporating new sampling schemes, such as wobbling and dichotomic sampling. The ring diameter is one of the key parameters in ring system construction. If constant crystal size and spacing are maintained, the advantages of an increased diameter are as follows: improvement of resolution uniformity, decrease in the number of random coincidences, more scatter rejection, and increase in the number of view angles. On the other hand, the expected disadvantages of an increased diameter include decreased sensitivity, resolution degradation due to angular uncertainty, and increased overall system costs. Ring diameter ranging from 45 to 65 cm appears to be suitable for brain scanners, while 70–90 cm is more suitable for body scanners.

v. Time-of-flight system. Ideally, if one can detect the exact difference in the flight times of two annihilation photons, the exact position of annihilation can be located, making possible a direct mapping of activity. The TOF concept in PET appears to be useful in low-resolution PET systems, where additional TOF information can help to enhance resolution and improve the signal-to-noise ratio of the image. To obtain any significant improvement, however, the time resolution should be substantially less than 500 ps. This is usually difficult to achieve by means of existing detector and electronic technology. The development of a few fast scintillation crystals, such as CsF and BaF_2, is of interest however.

vi. Spherical-pet system. For increasing overall sensitivity and direct volume imaging capability, the spherical shape of the PET system appears to be the most effective choice. The first extensive S-PET design and preliminary study were carried out by Cho *et al.* at Columbia University. The S-PET design was initially intended for high-resolution imaging, that is, resolution of $\sim$3 mm FWHM. To support this high resolution, it is imperative to maximize system sensitivity without impairing the true-to-random as well as scatter coincidence ratios. Therefore, spherical geometry and slice collimators focused on the system center have been incorporated. High stopping power detectors, such as BGO (with a detector width as narrow as 4 to 5 mm), have been proposed for this purpose. Special PMTs such as the rectangular PMT developed by Hamamatsu (the R2404) and new dichotomic sampling scheme have also been incorporated. In the field of image reconstruction, a TTR algorithm (discussed in Section II.A) has also been utilized. The design concept is currently under extensive evaluation from both the theoretical and experimental points of view.

2. Single Photon Emission Computerized Tomography

Although SPECT was first envisioned before the PET scanner, it is generally considered to be inferior to PET in a few critical aspects of imaging, including the capability of attenuation correction and ultimate resolution attainable with conventional collimators. Interest in SPECT has been renewed, however, due to its simplicity and availability.

In SPECT, any radioisotope that emits γ rays can be used. In contrast to annihilation photons, these γ rays are emitted as single individual photons. Isotopes common in SPECT imaging include ^{99m}Tc, ^{125}I, and ^{131}I. These are the radionuclides most often used in nuclear medicine.

Because of the nature of the isotopes used in SPECT [i.e., decay by the emission of (single) γ photons], a device for defining ray direction—the collimator—is required. This collimation process eliminates most of the γ rays emitted into 4π space. Unlike PET, this physical collimation inherently limits sensitivity.

The first rotating single-photon tomographic imaging system was developed by Kuhl and Edwards in 1963, without the benefit of modern tomographic image reconstruction techniques. Since the development of X-ray CT and PET, the instrumentation for SPECT has been extensively developed. Although most of the developments are based on the rotating Auger camera arrangements (camera-based approaches), a few systems utilize discrete detector arrays (scanner-based approaches).

The resolution of a SPECT image is directly related to the response of collimators employed and the sampling intervals in linear and angular directions. The response of the collimators determines resolution as well as system sensitivity and other factors, such as uniformity. The resolution of SPECT is poorer than that of PET, mainly due to collimation and partly due to the limited number of photons that can be collected. Axial resolution is especially dependent on the properties of collimation and depth variant.

As research into SPECT continues as expected, resolution and sensitivity will be improved through the application of new collimators and the use of new geometry, such as the triangular shape approach proposed by Lim *et al.*

C. Nuclear Magnetic Resonance Computerized Tomography

NMR CT is a 3-D imaging system that uses the NMR phenomenon as an imaging tool. Magnetic resonance is a phenomenon found in magnetic systems that possess both magnetic moment and angular momentum. The term *resonance* implies that we are in tune with a natural frequency

of the magnetic system. In this case, it corresponds to the frequency of gyroscopic precession of the magnetic moment of nuclei in an external static magnetic field. Because the magnetic resonance frequencies fall typically into the radio frequency range for nuclear spins, we often use the term *radio frequency* in referring to NMR. In NMR CT, we select a region from samples and obtain spatial distributions of nuclear spins through the application of deliberately added spatial field gradients and RF signals. Thus, the cross-sectional images of an object are obtained.

The advantages of NMR CT are its nonhazardous nature, its high-resolution capability, its potential for chemically specific imaging, its capability of obtaining anatomical cross-sectional images in any direction, and its high tissue discrimination capability (high-contrast resolution among different tissues). Although it has some minor disadvantages, such as its inherently long data acquisition time due to spin–lattice relaxation time and low signal-to-noise ratio, due to its many advantages the NMR CT is rapidly becoming a major diagnostic tool. NMR CT is overcoming the problem of time-consuming data acquisition through the development of new high-speed imaging methods and is surmounting the problem of low signal-to-noise ratio through the use of high-field magnets (superconducting magnets) with a magnetic field as high as 2.0 tesla (T) (1 T is equal to 10 kG) or more.

In the early 1970s, both Lauterbur and Damadian showed that NMR spectroscopic techniques could be applied to imaging the human body and demonstrated that these techniques might eventually be applicable to diagnostic imaging techniques similar to those of X-ray CT. In 1978, Andrew demonstrated the very high resolution capability of NMR by obtaining a fine and detailed image of the submillimeter septum of a small lemon. Moore and Holland soon followed with images of the human head, demonstrating definitively the potential of NMR tomography in diagnostic imaging.

NMR tomographic images can be formed by direct mapping, projection reconstruction, or Fourier imaging. To date, two kinds of potentially useful imaging methods—direct Fourier imaging and projection reconstruction—are most widely used.

An interesting aspect of NMR imaging is its diversity in image formation, data collection, and reconstruction. Many different imaging and data processing methods are presently known and will be discussed from various points of view, such as imaging time, available field gradients, pulse strength, speed, signal-to-noise ratio, and artifacts associated with the restored object image.

Great advances have also been made in the area of instrumentation for whole-body NMR tomography. The formation of an NMR CT system requires a magnet, gradient coils, RF coils, computer and peripherals, and associated electronics. A main-field strength of 0.5 to 20 kG is used with a field gradient of 0.01 to 1 G/cm, formed by x, y, and z gradient coils. In the case of proton imaging with these strengths of magnetic field and gradients, the required RF range is approximately 2 to 85 MHz. The expected resolution in a conventional NMR imaging system depends on the field homogeneity and the available strength of field gradients.

Two main types of imaging methods—direct Fourier imaging (Kumar–Welti–Ernst method) and projection reconstruction—and the possibility of parameter imaging will be discussed in detail here. Hardware configurations and the related statistical aspects of image quality and imaging times will also be discussed briefly.

A typical NMR tomographic imaging system designed for human imaging is depicted in Fig. 13a. In this figure, a split-solenoidal type of magnet NMR CT system is shown. The sample is surrounded by an RF coil (Fig. 13b) and a gradient coil set (Fig. 13c, d, and e). The configurations of the magnet and the gradient coils may differ depending on the design scheme, but the basic concept will be similar for the majority of NMR imaging systems in the future.

1. Principles of Nuclear Magnetic Resonance Tomography

a. Nuclear magnetic resonance physics. Since NMR was discovered in 1946, it has become an indispensable analytical method and tool in chemistry and physics. Although the basic physical properties of NMR are well known and information on NMR can readily be found elsewhere, we shall discuss a few topics necessary for understanding NMR as an imaging tool.

All materials, whether organic or inorganic, contain nuclei, which are protons, neutrons, or a combination of both. Nuclei that contain an odd number of protons, neutrons, or both in combination possess a nuclear "spin" and a "magnetic moment." This situation is equivalent to the aggregation of many small magnets. In the real world many materials are composed of several nuclei and the most common nuclei with magnetic moment are ^{1}H, ^{2}H, ^{7}Li, ^{13}C, ^{23}Na, ^{31}P, and ^{127}I. Although some materials are composed of nuclei with an even number of protons and neutrons that possess no spin or magnetic moment, they often contain some nuclei with an odd number of protons or neutrons. Therefore, they are subjects of NMR imaging. For this reason, NMR is practically applicable to most solid- and liquid-phase materials. Among the many hundreds of known stable nuclei, more than 100 possess spin and magnetic moment.

When a given material is placed in a magnetic field, some of its randomly oriented nuclei experience external

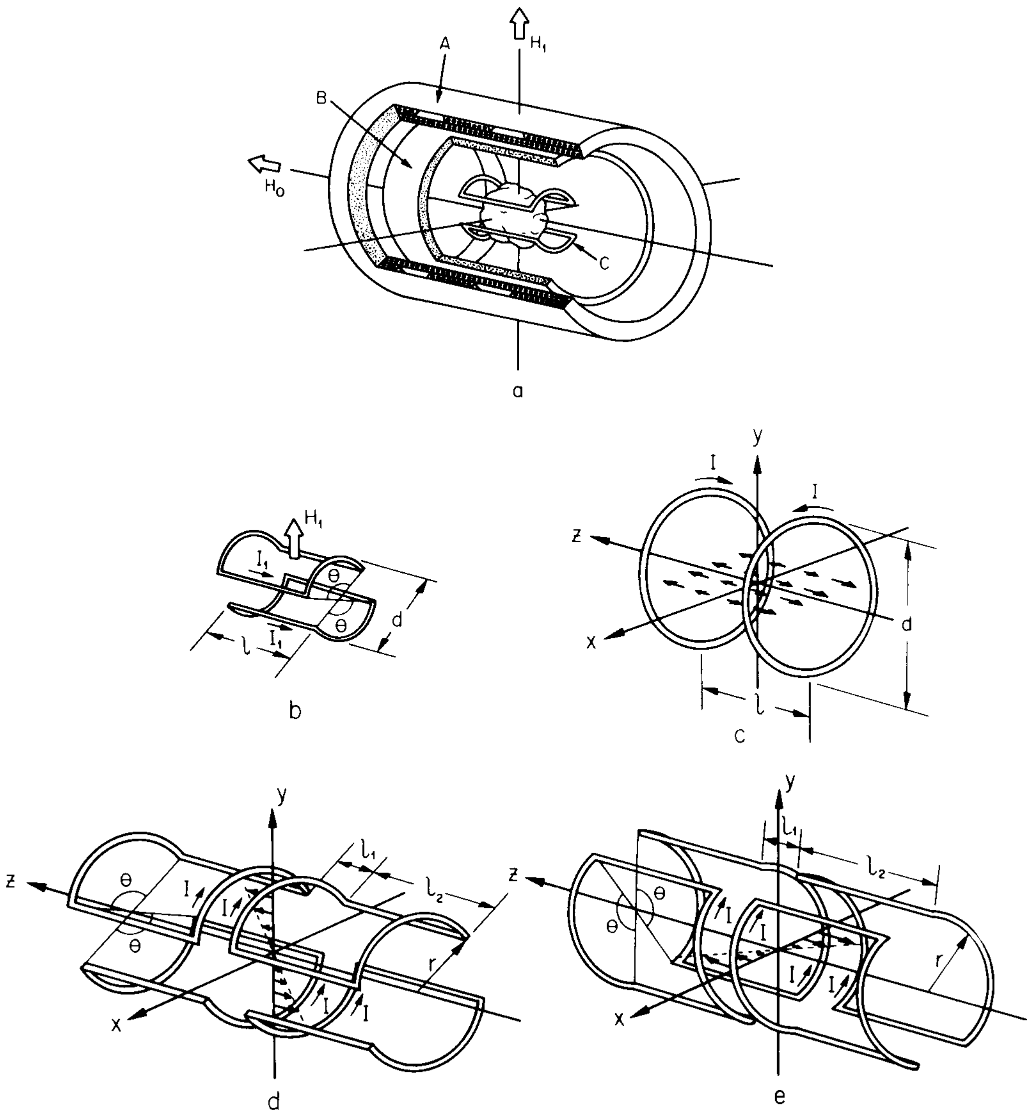


FIGURE 13 Sketch of an NMR tomograph. (a) Physical arrangement of the central part of an NMR CT system. A, Main magnet of the split-solenoidal type used in many superconducting magnets; B, gradient coil set; C, RF coil in the saddle type. (b) Saddle-shaped RF coil. Although the optimized shape for the homogeneous RF field is $\theta = 120°$ and $l/d = 2$, the shape can be changed slightly to accommodate the different sample shape or number of coil turns. (c) z-directional gradient coil. The change of the magnetic field in the z-direction is indicated by the length of the arrows. This coil is called the Maxwell pair. (d) y-directional gradient coil; often known as the Golay coil. (e) x-directional gradient coil; also of the Golay type. In (c), (d), and (e), the optimal coil shapes are $\theta = 120°$, $l/d = \sqrt{3}/2$, $l_1 = 0.78r$, and $l_2 = 2.13r$.

magnetic torque, which tends to align the nuclei in both parallel and antiparallel directions to the applied magnetic field. The fraction of magnetized nuclei in the direction parallel to the applied magnetic field is limited by thermal agitation. Therefore, it is also limited by the temperature and main magnetic field strength. Because this fraction is relatively small at room temperature, it has been a limiting factor in the sensitivity of NMR imaging. The spinning nucleus responds to the external magnetic field like a gyroscope precessing around the direction of the gravitational field. The rotating or precessional frequency of the spins, called the Larmor precession frequency, is proportional to the magnetic field strength.

Another important phenomenon of NMR is the creation of an energy "absorption state" (from a statistical point of view) by the applied external magnetic field. The proton has an intrinsic angular momentum or spin of $\hbar/2$, where $\hbar$ is Planck's constant divided by 2π. When

proton nuclei are placed in a magnetic field, the nuclei are in two energy states $+\mu H_0$ (antiparallel) and $-\mu H_0$ (parallel to the static magnetic field H_0), where μ and H_0 are the nuclear magnetic moment and applied magnetic field, respectively (Zeeman splitting). Because at thermal equilibrium the distribution of spins in energy states follows the Boltzmann law, the lower energy state has a larger population of spins than the higher energy state. For those nuclei or protons at the $-\mu H_0$ energy state or the lower energy state whose magnetic moments are parallel to H_0, the irradiation of external electromagnetic radiation of energy E equivalent to $2\mu H_0$ tends to excite protons at the $-\mu H_0$ energy states up to the higher energy $+\mu H_0$ state. This energy is given in the form of RF magnetic field H_1. The excited protons then tend to return to their low-energy state, producing the FID signal.

Two relaxation mechanisms are associated with these excited nuclear spins: transverse, or spin–spin, relaxation and longitudinal, or spin–lattice, relaxation. It is interesting that both these relaxation times (T_1 and T_2) are sensitive to the molecular structures and environments surrounding the nuclei. For example, the mean T_1 values of normal tissues and of many malignant tissues differ substantially from one another, allowing us to differentiate malignant tissues from normal tissues in many cases (Fig. 14c). A similar tendency is observed for T_2 values. The imaging capabilities of these two important parameters, T_1 and T_2, together with the spin densities of the objects, make NMR imaging a unique, versatile, and powerful technique in diagnostic imaging. Let us now review a few of the fundamental processes involved in NMR tomographic imaging.

Although many features of NMR phenomena can be understood only by quantum mechanical considerations, a number of properties are more easily visualized by means of a classical treatment. Let us consider a magnetic moment $\boldsymbol{\mu}$ in the presence of a magnetic field $\mathbf{H}_0$. Figure 15a depicts the precession of proton spins in two energy states. All moments precess about $\mathbf{H}_0$ at the same frequency, but without phase coherence in the x, y plane. Since the Boltzmann distribution favors the lower energy state, at equilibrium there are more nuclei aligned in the direction of $\mathbf{H}_0$. The net magnetization vector $\mathbf{M}_0$, which is the vector sum of $\boldsymbol{\mu}$'s, is oriented along the z axis (Fig. 15b). When the net magnetization vector $\mathbf{M}_0$ is at an angle θ to $\mathbf{H}_0$, the net energy of the system is

$$E = -\mathbf{M}_0 \cdot \mathbf{H}_0 = -M_0 H_0 \cos\theta. \quad (36)$$

Note that the spin system is in its lowest energy state when $\mathbf{M}_0$ is parallel to $\mathbf{H}_0$. The magnitude of the net magnetization at equilibrium is given by

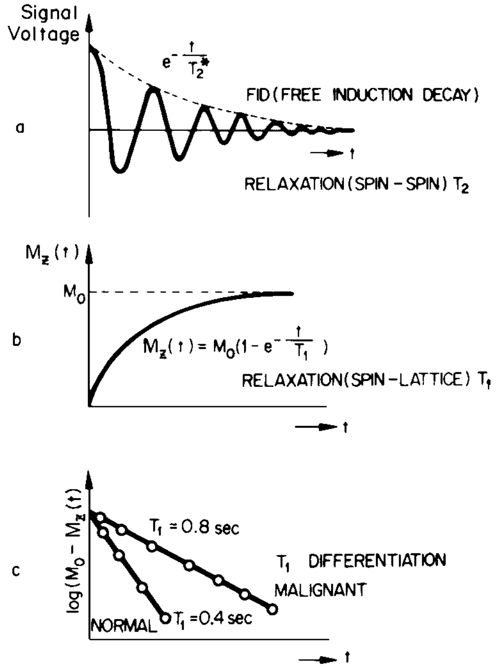


FIGURE 14 Spin–relaxation mechanisms. (a) FID signals obtained indicate a modulated decaying signal. The decay-time constant is T_2^*, which is also the effective spin–spin relaxation time. (b) Spins also decay by dissipating energy to the surroundings. This energy-dissipation mechanism is known as spin–lattice relaxation and is usually slow and decays with time constant T_1. (c) These relaxation mechanisms are expected to be used in discriminating malignant and normal tissues in NMR tomography.

$$M_0 = N(-\gamma h)^2 H_0 I(I+1)/3kT_0, \quad (37)$$

where N is the number of spins, γ is the gyromagnetic ratio, I is the spin quantum number, k is the Boltzmann constant, and T_0 is the object temperature. The signal strength, which is proportional to M_0, can be increased

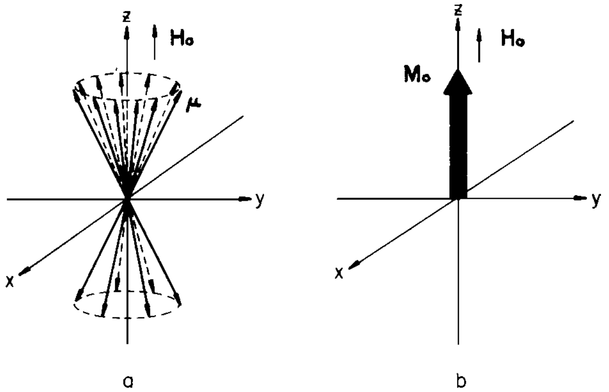


FIGURE 15 Spins in a magnetic field $\mathbf{H}_0$. (a) Spins precess about $\mathbf{H}_0$ in two energy states. More spins are aligned in the direction of $\mathbf{H}_0$. (b) Net spin magnetization vector $\mathbf{M}_0$ is given by $\mathbf{M}_0 = \Sigma\boldsymbol{\mu}$. Note that, at thermal equilibrium, $\mathbf{M}_0$ is along $\mathbf{H}_0$.

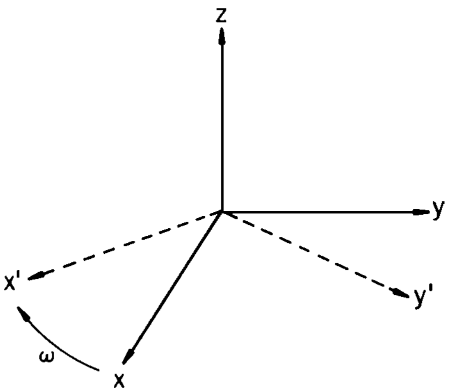


FIGURE 16 Rotating frame of reference. The coordinates (x', y', z') rotate about the z axis with the angular frequency ω. The rotating coordinates are related to the fixed coordinates (x, y, z) as $x' = x\cos\omega t + y\sin\omega t$; $y' = -x\sin\omega t + y\cos\omega t$; $z' = z$.

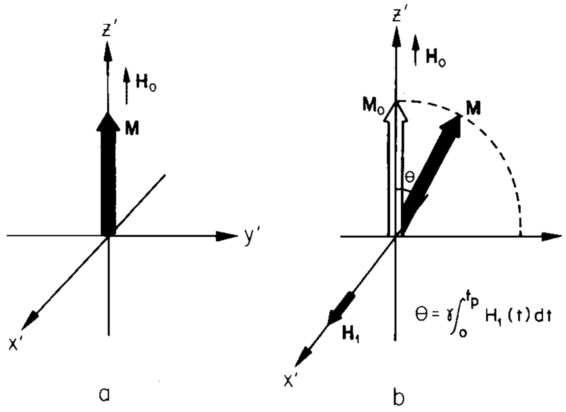


FIGURE 17 Spin magnetization in the rotating frame with and without RF pulse. (a) Spin in the absence of RF pulse. (b) Spin flip with an application of the RF field $\mathbf{H}_0$. The flipping angle θ of the magnetization is given by $\theta = \gamma\int_0^{t_p} H_1(t)dt$, where $H_1(t)$ is the time-varying RF field intensity and t_p the length of the RF pulse. The angle θ is usually set to 90° or 180°.

by increasing the field strength H_0. Lowering T_0 would also improve the equilibrium magnetization.

Spin precession can be observed by solving the differential equation of motion (the Bloch equation), given by

$$d\mathbf{M}_0/dt = \gamma\mathbf{M}_0 \times \mathbf{H}_0. \quad (38)$$

The resulting precession of spin follows the Larmor precession frequency,

$$\omega_0 = -\gamma\mathbf{H}_0, \quad (39)$$

which is unique to each nucleus. The minus sign indicates the clockwise precession for positive γ.

In visualizing the motion of the magnetization, it is convenient to use a rotating frame of reference. Let us introduce a set of Cartesian coordinates (x', y', z') rotating about $\mathbf{H}_0$ at an angular frequency ω (Fig. 16). The magnetic field associated with this frame is called the effective magnetic field, which is given by

$$\mathbf{H}_{\text{eff}} = \mathbf{H} + \omega/\gamma. \quad (40)$$

In the absence of an RF field, $\mathbf{H} = \mathbf{H}_0$. At resonance, therefore, the fictitious field ω/γ exactly cancels $\mathbf{H}$, and $\mathbf{H}_{\text{eff}}$ becomes zero. When the static magnetic field is in the z direction and the RF field $\mathbf{H}_1$ is applied along the x' direction (in other words, $\mathbf{H}_1$ is rotating clockwise in the x–y plane), the total magnetic field $\mathbf{H}$ is

$$\mathbf{H} = H_0\hat{z} + H_1(\hat{x}\cos\omega t + \hat{y}\sin\omega t), \quad (41)$$

where $\hat{x}$, $\hat{y}$, and $\hat{z}$ represent the unit vectors in the x, y, and z directions, respectively. If we insert Eq. (41) into Eq. (40), $\mathbf{H}_{\text{eff}}$ becomes

$$\mathbf{H}_{\text{eff}} = \left(H_0 - \frac{\omega}{\gamma}\right)\hat{z}' + H_1\hat{x}', \quad (42)$$

where $\hat{x}'$ and $\hat{z}'$ are the unit vectors in the x' and z' directions, respectively. At resonance ($\omega = \omega_0$), Eq. (42) can be expressed as $\mathbf{H}_{\text{eff}} = H_1\hat{x}'$. In this case, in a rotating frame the only magnetic field is in the x' direction, and $\mathbf{M}$ precesses around the x' axis or $\mathbf{H}_1$ with frequency γH_1. For a general time-varying RF field $H_1(t)$, the flipping angle is given by

$$\theta = \gamma\int_0^{t_p} H_1(t)\,dt, \quad (43)$$

where t_p is the RF pulse duration. The application of an RF pulse, which tips the magnetization $\mathbf{M}$ into the x–y plane, causes the excitation of the spin system (Fig. 17). When H_1 is applied along the x' axis for a pulse period t_p, the spin rotates or flips through an angle θ from the z' axis toward the y' axis. In general, θ is set at $\pi/2$ or π, depending on the mode of excitation and the type of NMR experiments. In the simplest case, $\theta = \pi/2$ is used to observe the maximum transverse component of magnetization.

After H_1 is turned off, the rotating magnetization induces a current into the pickup coil surrounding the object. The magnetization then relaxes, through neighboring spins and environment, to its thermal equilibrium, so that the spins realign with the original H_0 field direction. On the other hand, the transverse component of magnetization, which is related to the entropy of the system, decays through the spin–spin interaction and dephases.

In addition to the inherent spin–spin relaxation, there are other dephasing effects, such as the magnetic field inhomogeneity and field gradients. In NMR imaging, magnetic field gradients are deliberately added to resolve the spatial distribution of spin density. In fact, they produce shifts in the Larmor frequencies throughout the sample, resulting in a phase incoherency that eventually makes the composite sinusoidal signal decay more rapidly than the inherent transverse relaxation time T_2. This effective

transverse relaxation time resulting from field inhomogeneity alone is expressed as

$$1/T_2^* = 1/T_2 + \gamma \Delta H/2, \qquad (44)$$

where ΔH is the field inhomogeneity, that is, the maximum deviation of magnetic field over the object region. When a field gradient is added to resolve the spatial distribution of spin density, T_2^* is further reduced to T_2^{**}, as given by

$$1/T_2^{**} = 1/T_2^* + \gamma G R, \qquad (45)$$

where G (in gauss per centimeter) is the gradient field strength and R (in centimeters) is the object diameter. The composite sinusoidal signal decaying with an effective transverse relaxation time T_2^{**} is then detected with a phase-sensitive detector. The results are similar to a decaying demodulated AM signal, as shown in Fig. 14a.

Concurrently, longitudinal or spin–lattice relaxation forces the spins to realign in the H_0 (or z) direction because it is the lowest energy state or thermal equilibrium state. Because it involves energy dissipation through the lattice, the longitudinal relaxation time T_1 is usually longer than T_2 and is related to the z component of magnetization, as stated in the following (see Fig. 14b),

$$M_z = M_0 \left[1 - \left(1 - \frac{M_z'}{M_0} \right) \exp(-t/T_1) \right], \qquad (46)$$

where M_z' is the z component of magnetization at the starting time of relaxation.

The two relaxation processes work simultaneously and vary greatly depending on the characteristics of the material. In the case of tissue in field strength of 1 to 20 kG, for instance, T_1 and T_2 are of the order of 0.5 s and 50 ms, respectively; T_1 is usually larger than T_2. In Fig. 18, sequential pictures of the relaxation processes are shown.

In general, T_1, T_2, T_2^*, and T_2^{**} have the following relationship:

$$T_2^{**} \le T_2^* \le T_2 \le T_1. \qquad (47)$$

When the two relaxation mechanisms are considered, the Bloch equation can be written

$$\begin{aligned} \frac{dM_z}{dt} &= \gamma(\mathbf{M}_0 \times \mathbf{H}_0)_z - \frac{M_z - M_0}{T_1} \\ \frac{dM_{xy}}{dt} &= \gamma(\mathbf{M}_0 \times \mathbf{H}_0)_{xy} - \frac{M_{xy}}{T_2}, \end{aligned} \qquad (48)$$

where $(\cdot)_z$ and $(\cdot)_{xy}$ represent z and x–y components, respectively. Equation (48) indicates that the magnetization components M_z and M_{xy} are independently related to the relaxation times T_1 and T_2.

Several forms of spin-echo techniques play a central and essential role in data acquisition for NMR imaging. The two basic forms of the spin-echo technique applicable to NMR imaging are the Hahn spin-echo technique and the Carr–Purcell Meiboom–Gill (CPMG) technique. In the Hahn spin-echo technique, a 90° RF pulse is applied to the direction of the x' axis, and then the magnetization vector $\mathbf{M}$ rotates to the y' axis. The spin magnetizations then dephase over time, due to field inhomogeneity or added field gradients. A subsequent 180° pulse applied along the x' axis rotates the spins around the x' axis (Fig. 19b). The spin magnetizations now continue to precess but begin to rephase. This process is equivalent to a focusing or regrowing process of the FID signal at twice the dephasing time. At this point, all the spins are completely rephased along the $-y'$ axis, but the magnetizations have now decayed by T_2 relaxation.

In the CPMG method, a 180° pulse is applied along the y' axis instead of the x' axis, so that the spin flips around

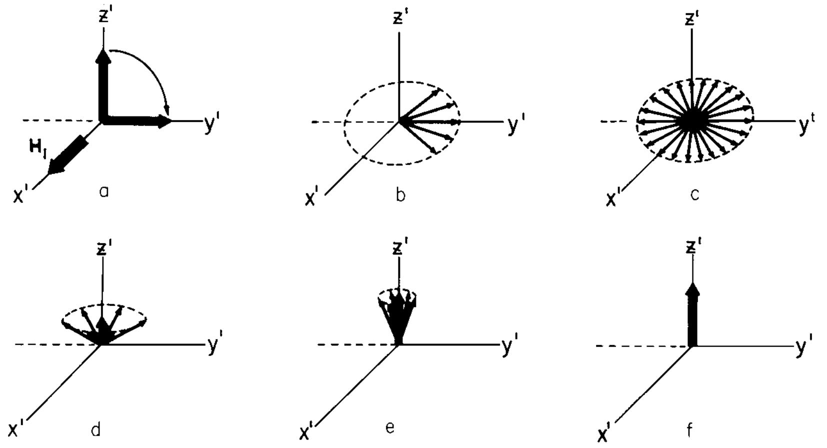


FIGURE 18 Sequential illustrations of the spin relaxation processes. (a) Spins are flipped by the RF pulse $\mathbf{H}_1$; (b) spins are dephased due to the spin–spin relaxation and field inhomogeneity; (c) FID signal decays to zero as the spins lose phase coherence; (d), (e), (f) spins relax to the original equilibrium state by the spin–lattice relaxation process.

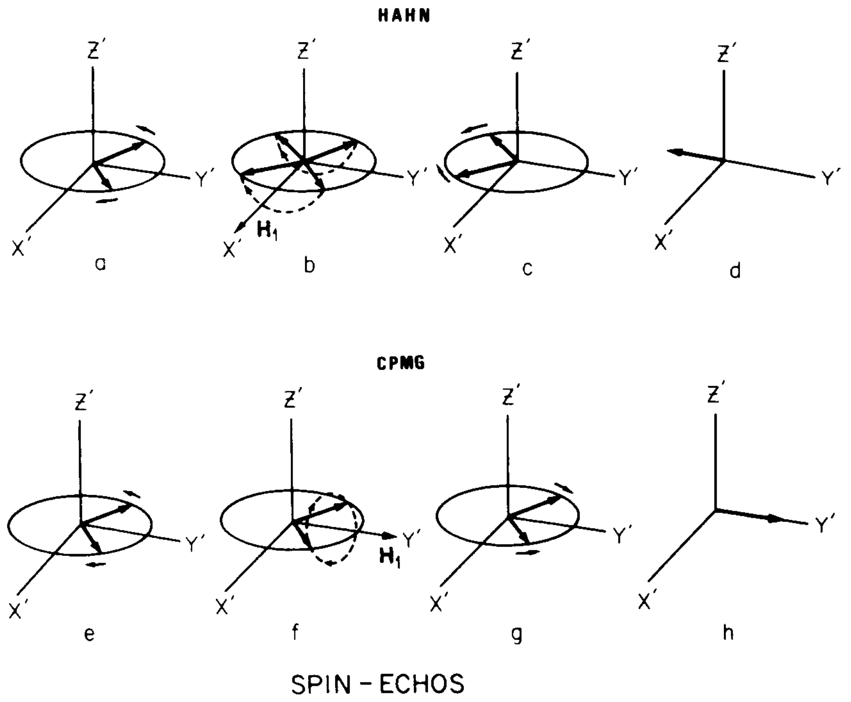


FIGURE 19 Hahn spin echo: (a) Spin magnetizations dephase after 90° rotating by RF pulse; (b) 180° pulse is applied along x' axis; (c) the spins are being refocused; (d) spin echo is generated along $-y'$ axis. Carr–Purcell and Meiboom–Gill spin echo: (e) Spins are dephased after 90° rotating by RF pulse; (f) 180° pulse is applied along y' axis; (g) the spins are being refocused; (h) spin echo is generated along y' axis.

the y' axis (Fig. 19f). Both techniques are actively used in all phases of NMR imaging to reduce several adverse effects that arise in actual data collection, such as field inhomogeneity and the effects of the gradient pulse rise time.

b. Basic theory of nuclear magnetic resonance tomography. Conventional NMR chemistry requires a magnetic field of extreme homogeneity, in other words as uniform as possible, to reduce the frequency shift effect caused by the spatially dependent field variations. In Fourier NMR imaging, however, a field gradient or set of gradients is deliberately added to resolve the spatial distribution of spins into Fourier domain components. The basic form of signal obtained from 3-D Fourier transform NMR, which is known as FID, is expressed as

$$s(t) = M_0 \iiint_{-\infty}^{\infty} f(x, y, z) \exp\left\{ i\gamma \int_0^t [xG_x(t') + yG_y(t') + zG_z(t')]\, dt' \right\} dx\, dy\, dz \qquad (49)$$

where $f(x, y, z)$ is the 3-D spin density distribution and $G_x(t)$, $G_y(t)$, and $G_z(t)$ are the time-dependent field gradients along the x, y, and z axes, respectively. In Eq. (49) the effects of T_1 and T_2 relaxation times are not included; they will be discussed in a later section. The generated FID is, in effect, a Fourier transform-domain representation of the spin density distribution. From this fundamental 3-D equation, many equations for the imaging algorithms described below can be derived.

2. Image Formation Algorithms

One of the interesting aspects of NMR imaging is that the NMR images can be formed by many different procedures, some of which are described in the following text. The discussion is limited to the techniques used most often, however.

In NMR imaging, data acquisition pulse sequences play an important role and are intimately related to the image reconstruction algorithms employed. This description of the mathematical formulations is based on the basic pulse sequence, which uses only 90° RF pulses, although in a real imaging situation the spin-echo techniques that use additional 180° RF pulses are more common.

a. Direct Fourier imaging.

i. Algorithm for direct Fourier imaging technique. This direct Fourier imaging (DFI) method was first proposed by Kumar, Welti, and Ernst (KWE). In this case,

imaging can proceed through the total 3-D excitation of an object in series of time sequences. The result of the 3-D Fourier transform of those data is considered to be the 3-D spin density function or image.

In this DFI or 3-D KWE procedure, three orthogonal field gradients, G_x, G_y, and G_z, are applied in sequence after 90° RF excitation pulse at $t = 0$. The z components of the local magnetic fields are given as

$$H_z(x, y, t) = \begin{cases} H_0 + G_z z, & \text{for } 0 < t < t_z \\ H_0 + G_y y, & \text{for } t_z < t < t_z + t_y \\ H_0 + G_x x, & \text{for } t_z + t_y < t < t_z + t_y + t_x, \end{cases} \tag{50}$$

where each timescale t_z and t_y is varied according to preassigned sequences, that is, $t_z \simeq t_a$, $t_y \simeq 0 \sim t_a$, where t_a is the optimal observation time of FID. The FID signal is sampled when the x gradient is applied, namely, during the t_x period. The sampled FID signal reflects the previous application of the z and y gradients by retaining the phase change caused by those gradients. For this reason, this imaging scheme is often called the phase-encoding method. The series of FID signals obtained with the various t_z's and t_y's then form a full 3-D FID signal set sufficient for reconstruction of the spin density image of the entire volume.

The observed FID signal $s(t_x, t_y, t_z)$ is (neglecting the relaxation processes) expressed as

$$s(t_x, t_y, t_z) = M_0 \iiint_{-\infty}^{\infty} f(x, y, z) \exp[i\gamma(G_x x t_x + G_y y t_y + G_z z t_z)]\, dz\, dy\, dx. \tag{51}$$

Fourier transform of Eq. (51) results in spatial spin density function as

$$\tilde{f}(\omega_x, \omega_y, \omega_z) = \iiint_{-\infty}^{\infty} s(t) \exp[-i(\omega_x t_x + \omega_y t_y + \omega_z t_z)]\, dt_z\, dt_y\, dt_x. \tag{52}$$

The reconstructed image $\tilde{f}(\omega_x, \omega_y, \omega_z)$ is related to $f(x, y, z)$ as

$$\tilde{f}(\omega_x, \omega_y, \omega_z) = k f(\gamma x G_x, \gamma y G_y, \gamma z G_z), \tag{53}$$

where k is a constant.

Figure 20 shows RF and gradient pulse sequences of conventional DFI using spin echo for one-slice 2-D imaging. In this scheme, although the whole volume of an object is excited by the 90° RF pulse, only the spins in a designated slice are rephased to form an echo through the application of a narrowband 180° RF pulse and z-directional selection gradient. Note that the x-directional phase encoding is achieved by varying the amplitude of the x gradient instead of varying the time interval.

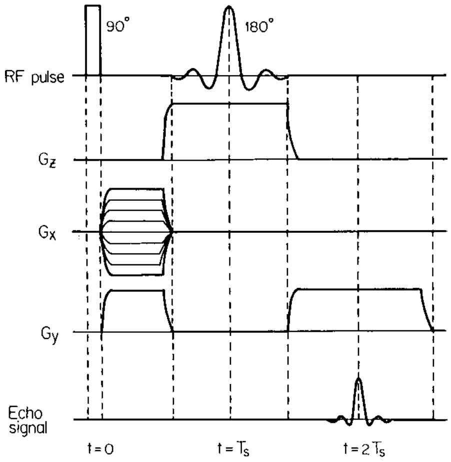


FIGURE 20 Imaging sequences of 2-D direct Fourier imaging. The slice in the z direction is selected and spin echo is used. While G_y remains constant, the intensity of G_x is varied for phase encoding. The purpose of the first part of the G_y gradient pulse is to dephase spins in the object after the nonselective 90° RF pulse. Only the spins in a designated slice are rephased by the selective 180° pulse and form the spin echo at $t = 2T_S$ on the second constant G_y gradient. The spin-echo signal is collected for image reconstruction.

ii. Time-multiplexed multislice imaging method. By using consecutive excitations of the pulse sequence shown in Fig. 20 within a suitable repetition time period, it is possible to obtain several images simply through the use of the remaining time, provided that the sum of data acquisition time of each slice is smaller than the repetition time. For example, a data acquisition time of less than 100 ms with a repetition time of 300 ms would allow three consecutive slice imagings without T_1 relaxation constraints. This multislice imaging method is a frequently used NMR imaging technique. It can also be applied to other 2-D imaging schemes, such as the line-integral projection reconstruction (LPR) technique, which is described in the following section.

b. Line-integral projection reconstruction.

i. Basic principles. Projection reconstruction using 2-D and 3-D image reconstruction algorithms is well known, especially in the areas of X-ray CT and radionuclide emission tomography, as previously discussed. Although the image can be reconstructed in several different ways, the basic forms of data collection are similar. Line-integral projection data are obtained in angular steps, by rotating the object a total of either 180° or 360°. The most familiar and convenient way to reconstruct 2-D or 3-D images is through the Fourier convolution method, which can

be summarized as follows. The reconstructed 2-D image $f(x, y)$ is given by

$$f(x, y) = \int_0^{\pi} [p_\phi(x') * h(x')]\, d\phi, \qquad (54)$$

where $p_\phi(x')$ is projection data, $h(x')$ is the filter kernel that corrects $1/r$ blurring caused by circular symmetric linear superposition, and (x', y') is the coordinate system rotated by an angle ϕ from the original coordinates (x, y).

In Fourier transform NMR, the nuclear signal can be considered to be the inverse Fourier transform of the spatial domain spin density function. If a plane at $z = z_0$ is selected, then the FID at an angular view ϕ can be expressed as

$$s_\phi(t) = M_0 \iint_{-\infty}^{\infty} f(x', y'; z_0) \exp[i\gamma x' G_{x'} t]\, dy'\, dx'. \qquad (55)$$

Although the FID signal $s_\phi(t)$ appears in the time domain, it represents the Fourier domain projection data. Therefore, the projection data $p_\phi(x')$ are obtained through the Fourier transform of the FID signal as

$$p_\phi(x') = \mathcal{F}[s_\phi(t); t \rightarrow x']. \qquad (56)$$

The basic form of projection data obtainable in Fourier transform NMR is similar to the data obtained in X-ray CT. In Fig. 21, spin-echo signals or FIDs are obtained at different angular views through the application of the field gradient and RF excitation sequences. As a first step, all the spins in the sample are excited with a 90° RF pulse, and a subsequent 180° RF selects the slice. After spins are refocused they generate the spin-echo or FID signal (Fig. 21). After 180° or 360° rotation of projection with an appropriate step through the adjustment of the field gradients G_x and G_y, a complete projection data set sufficient for reconstruction of a slice at a given plane z_0 is obtained. At this point, 2-D image reconstruction can proceed according to Eq. (54); that is, each echo or FID signal $s_\phi(t)$ is Fourier-transformed, convolved with a filter kernel, and backprojected.

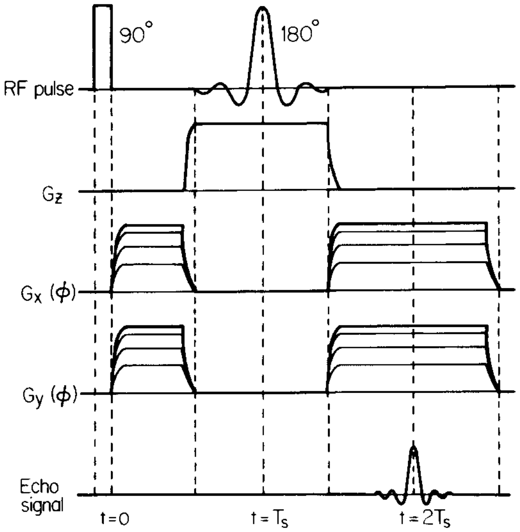


FIGURE 21 FID, RF pulse timing, and field gradient sequences for the basic single-slice line-integral projection reconstruction. Gradient pulse sequence: $G_x(\phi) = G\cos\phi$; $G_y(\phi) = G\sin\phi$; $0° \leq \phi < 180°$ or $0° \leq \phi < 360°$, where G is the maximum value of the reading gradient.

ii. Slice (plane)-encoded multislice LPR. The single-slice line-integral projection technique explained earlier can be extended to achieve multislice imaging through several encoding techniques, for example, the plane-encoding technique explained in the following paragraphs.

Let us assume that the number of planes is n. For the data set at a view ϕ_i, the same G_{xy} and G_z are applied n times, each with a different frequency composition of RF pulses. The RF pulses are specially tailored to assign desired phases to the designated slices. To obtain a complete set of view data corresponding to the n planes, the acquisition of data is repeated n times with differently composed RF pulses.

The key to this method lies in the encoding of signals according to the RF pulse sequence. A simple illustration of the encoding procedure using a coding matrix is as follows. Let the FIDs obtained at each 180° composite RF pulse sequence be $S^1_{\phi_0}(t), S^2_{\phi_0}(t), \ldots, S^n_{\phi_0}(t)$. Each FID is a composite of the line-integral projection sets, which include data from several planes at an angular view ϕ_0, that is, $s_{\phi_0 z_0}(t)$, $s_{\phi_0 z_1}(t)$, and so on. Therefore, composite FIDs, $S^1_{\phi_0}(t), S^2_{\phi_0}(t), \ldots, S^n_{\phi_0}(t)$ can be given as

$$\mathbf{S}_{\phi_0}(t) = \begin{bmatrix} S^1_{\phi_0}(t) \\ S^2_{\phi_0}(t) \\ \vdots \\ S^n_{\phi_0}(t) \end{bmatrix} = [H_n] \begin{bmatrix} s_{\phi_0 z_0}(t) \\ s_{\phi_0 z_1}(t) \\ \vdots \\ s_{\phi_0 z_{n-1}}(t) \end{bmatrix}. \qquad (57)$$

From Eq. (57), the desired FID signal $s_{\phi_i z_i}$, which corresponds to the FID of slice z_i, can be obtained through matrix inversion.

Examples of coding matrices include the Hadamard matrix and the Fourier matrix. The advantage of this method is the statistical improvement gained as a result of the increase in total scanning time.

c. Planar-integral projection reconstruction. In the planar-integral projection reconstruction (PPR) method, both nonselective broadband 90° and spin-echo 180° pulses are applied, thus providing FID data that

originate from the entire volume. The FID signal of the total volume planar-integral projection data in the z' direction can be expressed as

$$s(t) = M_0 \int_{-\infty}^{\infty} \left[\iint_{-\infty}^{\infty} f(x', y', z')\, dx'\, dy' \right] \times \exp[i\gamma(H_0 + z' G_{z'})t]\, dz', \quad (58)$$

where (x', y', z') are the rotated coordinates of (x, y, z) and the z' direction coincides with the projection direction. Through phase-sensitive detection, the FID signal obtained can be written

$$s_{\theta,\phi}(t) = M_0 \int_{-\infty}^{\infty} p_{\theta,\phi}(z') \exp(i\gamma z' G_{z'} t)\, dz', \quad (59)$$

where $p_{\theta,\phi}(z')$ is the planar-integral projection data with the angular view (θ, ϕ).

In Eq. (59), $s_{\theta,\phi}(t)$ represents the projection data in the Fourier domain. Therefore, the Fourier transform of $s_{\theta,\phi}(t)$ is spatial domain planar-integral projection data with which reconstruction can be performed (see the PPR algorithm in Section II.A).

d. Echo-planar imaging method. One of the inherent disadvantages of NMR CT is the long data acquisition time caused by constraints on spin–lattice relaxation time. The echo-planar imaging method is one of the faster imaging techniques; imaging time can be reduced to as short as 50 ms.

In this method, the FID data in the spatial frequency domain is acquired following one simple excitation through the application of the oscillating gradient during the signal reception. The image can be reconstructed through a simple 1-D Fourier transform of the obtained FID data. This method is limited by gradient driving power and speed. Nevertheless, the echo-planar technique itself or variations of the technique seem to be potential candidates for future fast imaging techniques in NMR CT.

3. Imaging Modes and Extraction of Nuclear Magnetic Resonance Parameters

The spin density $f(x, y, z)$ obtained by the various imaging methods described earlier is not a real spin density; it is weighted by T_1 or T_2 or both. Because T_1 or T_2 varies between normal and abnormal tissues, the image of spin density weighted by T_1 or T_2 has been found to be clinically useful. With this in mind, several attempts have been made to extract T_1 information, as well as spin density and T_2. The typical imaging modes and corresponding terminologies currently in use are described in the following subsections.

a. Imaging modes.

i. Saturation recovery imaging. The saturation recovery method involves simply repeating the pulse sequence at regular intervals T. The equations discussed previously are unchanged except for the replacement of $f(x, y, z)$ with $f'(x, y, z)$, which is expressed as

$$f'(x, y, z) = f(x, y, z)\{1 - \exp[-T/T_1(x, y, z)]\}. \quad (60)$$

Note that f' is now a function of both T_1 and f.

ii. Inversion recovery imaging. Inversion recovery is similar to saturation recovery, except that the 180° RF pulse precedes the 90° RF pulse with a time interval of T_1; $f'(x, y, z)$ is related to $f(x, y, z)$ as

$$f'(x, y, z) = f(x, y, z)\{1 - 2\exp[-T_1/T_1(x, y, z)]\}. \quad (61)$$

It is easy to see the increased dependency of the image on T_1 over that obtained in saturation recovery. This technique is often used for measuring T_1 values in tissues. Figure 22 shows the pulse sequences for 2-D inversion recovery Fourier imaging.

iii. Spin-echo imaging. Through the application of the 180° pulse following the first 90° pulse at $t = T_s$, spins are refocused at $t = 2T_s$ by the spin echo (Fig. 20). Although the spins are now refocused and coherent, the amplitude of FID decays exponentially with time constant T_2. The decayed spin density $f'(x, y, z)$ can be written

$$f'(x, y, z) = f(x, y, z)\exp[-2T_s/T_2(x, y, z)]. \quad (62)$$

As explained earlier, the image is now weighted by T_2 as well as by T_1. By setting the appropriate T_s values, images

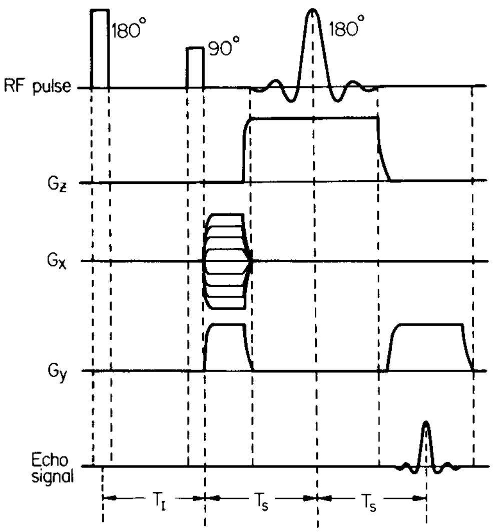


FIGURE 22 RF and gradient pulse sequences of inverson recovery direct Fourier imaging.

weighted mainly by T_2 can be obtained, provided that the repetition time is sufficiently large.

b. Parameter imaging methods. The capability of extracting many functional parameters is one of the most important advantages of NMR CT. Flow velocity, T_1, T_2, and chemical shift are some of the interesting parameters in NMR imaging that are discussed in this section.

i. T_1 (Spin–lattice relaxation time) and T_2 (Spin–spin relaxation time). The effects of T_1 and T_2 are closely related to the NMR imaging modes. In T_1 imaging, both the saturation recovery and inversion recovery modes can be used. By varying the recovery time and observing the resulting image intensity variation, one can deduce T_1 values. Similarly, by changing the echo time, that is, varying $2T_s$ in Eq. (62) for the spin-echo method, one can obtain several images differently weighted by T_2. From the images obtained with different echo times, T_2 values of each pixel can be calculated.

ii. Flow imaging. In NMR CT, one can also measure the flow or moving velocity of nuclear spins through observation of the FID signal. In the first attempt at flow velocity measurement two RF coils were used—one for the excitation of spins and the other for reception. In this experiment, surface RF coils were used to excite and receive the signal at known locations. If the maximum signal is received at Δt seconds after the excitation with the distance Δl between two RF coils, the velocity can be estimated by $\Delta l/\Delta t$.

Several flow imaging methods have been developed. Among these, two techniques relevant to general flow measurement will be discussed: one using density information and another using phase information.

The RF and gradient pulse scheme of flow imaging using the selective saturation method uses intensity information. In this scheme, the first 90° RF pulse and the homogeneity-spoiling gradients are used to saturate the spins in the selected slice for flow imaging. The 2-D Fourier imaging sequence for the same slice follows after Δt seconds to measure the signals originating from spins that flowed in from outside the slice, where spins were not saturated. From the density change observed for several different Δt's, the flow velocity in the selection gradient direction can be determined as $\Delta z/\Delta T$, where ΔT is the minimum Δt with the maximum spin density and Δz is the slice thickness.

In another variation of flow imaging, phase information is used to measure flow velocity. Since the pixel values of an image are usually extracted by taking the real part or absolute values of the image data in complex form, it is possible to use the phase information associated with each pixel data. Let us assume that a time-varying gradient $G_x(t)$ is applied to moving spins after RF excitation. The phase coding resulting from the time-varying gradient can be divided into two terms: the spatially coded term ϕ_s and the velocity-coded term ϕ_v, respectively. The sum appears as

$$\phi = \phi_s + \phi_v, \tag{63}$$

where

$$\phi_s = \gamma \int G_x(t) x \, dt$$

and

$$\phi_v = \gamma \int G_x(t) v t \, dt.$$

In Eq. (63), $G_x(t), x$, and v represent the time-dependent x gradient, the x coordinate of spins at $t = 0$, and the x-directional flow velocity of the moving spins, respectively. In the flow measurement, the flow coding gradient is applied in addition to the conventional RF and gradient pulse sequences, so that the phases on the final image are changed only as a result of flow velocity. Flow velocity can be determined from the calculated phase, which is coded according to the velocity of the spins. The unique advantage of this kind of flow velocity imaging method is the capability of multidirection flow imaging by simply applying the additional flow coding gradient in the desired direction. Figure 23 shows a typical gradient waveform for flow phase coding in the x direction, in which the spatially coded phase term is canceled so that $\phi_s = 0$, while the velocity-coded phase term remains $\phi_v \neq 0$. This technique, therefore, allows us to measure velocity by simply measuring the phase, which is now purely dependent on velocity.

iii. Chemical-shift imaging. Another important aspect of NMR CT is its spectroscopic imaging capability. Before NMR CT was proposed, NMR had been

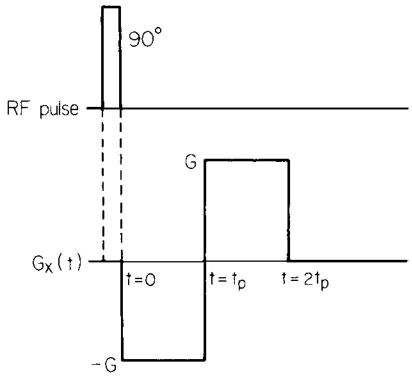


FIGURE 23 Gradient waveform for phase coding of flow velocity measurement. Note that the gradient pulsing shown effectively cancels out the spatial coding. The remaining velocity-coded phase ϕ_v can be written $\phi_v = \gamma t_p^2 G v$, where v is the flow speed.

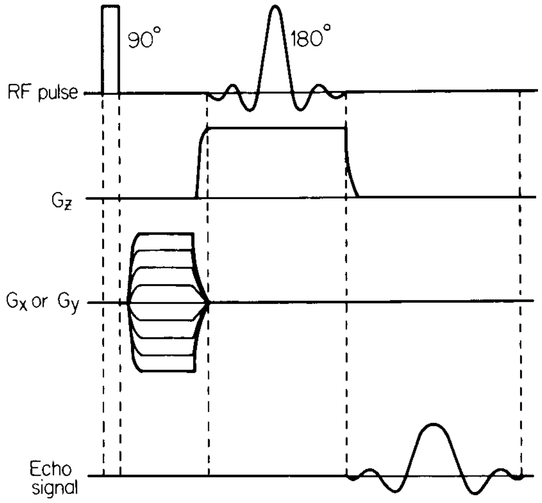


FIGURE 24 RF and gradient pulse sequences for 4-D chemical shift imaging using 2-D spatial codings. Note that the total number of coding steps for a slice is N^2.

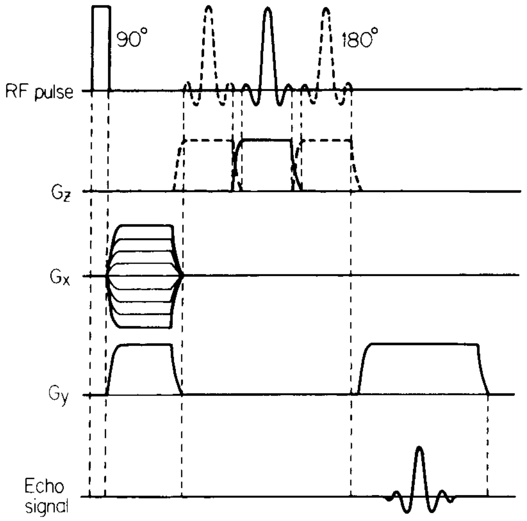


FIGURE 25 RF and gradient pulse sequences for the echo-time-encoded chemical-shift imaging. Note here that the total number of coding steps for a slice is N.

used primarily for chemical spectroscopy, in which the frequency spectrum of a specific kind of nuclei distributed over a few tens of parts per million of its Larmor precession frequency was obtained. The chemical shift was usually measured for homogeneous samples under the condition of uniform field.

In spectroscopic NMR imaging, however, the chemical spectrum for each pixel (chemical spectroscopic imaging) is to be measured neither for the homogeneous samples nor under the uniform field condition but with spatially varying gradient pulses.

A few chemical-shift imaging techniques have been proposed. An original spectroscopic imaging pulse sequence is shown in Fig. 24. The essence of this pulse scheme is the absence of the reading gradient during data acquisition. In this scheme, N^2 steps are required for a 2-D spectroscopic imaging of $N \times N$ matrix size image. In Fig. 25, a more generalized imaging sequence using echo-time encoding is shown. In this scheme, the 180° spin-echo RF pulses are applied several times, and corresponding FIDs are observed at each time. The notable difference between this scheme and the former is that here the spatial coding is identical to conventional 2-D imaging (i.e., gradient steps required are only N for $N \times N$ matrix size image), while in the former, the number of steps required is N^2. Also in the former, N determines the spectroscopic resolution. In the latter scheme, on the other hand, the number of RF time positions determines the spectroscopic resolution and, therefore, by varying the number of RF time positions, one can achieve the desired resolution. Often this step turns out to be much less than N in most *in vivo* spectroscopic imaging.

c. Other imaging methods. In addition to the imaging methods previously mentioned, there are several other imaging schemes of special form. One of these is gated or synchronized imaging, for an object that moves periodically. An example is the gated cardiac imaging of the human heart. In this case, the RF and the gradient pulse sequences are gated in synchronization with the ECG signals, and data in the different parts of the heart cycle are collected.

In the area of imaging methodology, rotating-frame zeugmatography should be noted. In this method, spatial-phase coding is achieved through the RF field gradient rather than spatial field gradients generated by the x-, y-, and z-directional gradient coils as in conventional NMR imaging. Although this method has some advantages, it is rarely used in imaging because of inherent difficulties, such as those found in the realization of the RF field gradient and associated RF coils.

4. System Configuration

The whole NMR CT system can be divided into three parts: the NMR section, the electronics, and the computer. The NMR section includes the main magnet, which provides the static main magnetic field; the gradient coil for generating magnetic field gradients; and the RF coil, which transmits and receives the RF signals. The electronic part includes a waveform synthesizer, a data acquisition component, and transmitter and receiver amplifiers. The computer performs image data processing and system control and finally displays the reconstructed images.

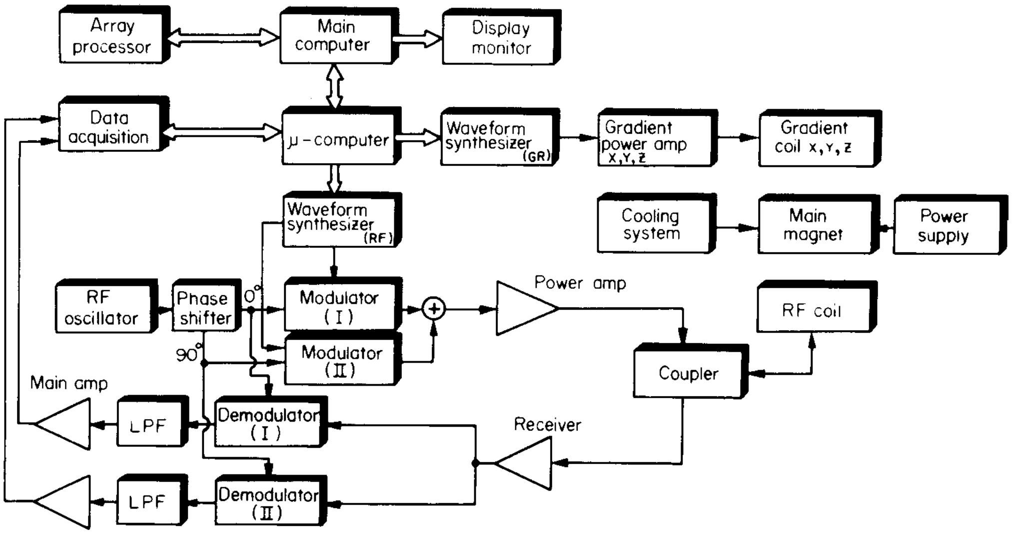


FIGURE 26 Block diagram of a typical NMR CT system.

A block diagram of a typical NMR CT system is illustrated in Fig. 26, which depicts the operation of each part. The main computer generates RF and gradient waveforms and reconstructs images after data acquisition. RF and gradient pulse waveforms generated in the main computer are transferred to a microcomputer and then to a waveform synthesizer, where data in digital form are converted to analog form. The gradient is applied to x-, y-, and z-gradient coils after being amplified in the gradient power amplifier. The RF waveform from the waveform synthesizer is modulated with the RF (reference) signal in the modulator, amplified through the power amplifier, and transferred to the RF coil via the coupler. The coupler circuit effectively switches on and off between the transmitting and receiving operations. The transmitted RF pulse excites nuclear spins in the sample. The nuclear signal induced on the RF coil by precessing spins is transferred to the receiver amplifier through the coupler. The amplified nuclear signal is demodulated with the RF reference signal and sent to the data acquisition part. Acquired nuclear signals (FIDs or echo signals) are transferred to the main computer via the microcomputer and are used for the reconstruction of the image. NMR CT systems often employ array processors for rapid image reconstruction. After reconstruction, the images are displayed on a cathode ray tube.

IV. RECENT DEVELOPMENTS

Among the many important recent developments in NMR imaging, the most notable one is the high-speed imaging using small flip angle gradient echo technique such as the steady-state free precession (SSFP) technique. This technique enables us to obtain high-quality images within a few seconds, compared to the conventional spin-echo technique, which usually requires an average of minutes or so for the imaging of a slice.

A typical pulse sequence for the fast gradient echo technique known as SSFP is illustrated in Fig. 27. The unique feature of the method, in addition to the speed advantage, is the potential of obtaining two characteristically different image data simultaneously, namely FID and echo. Here the FID image refers to the T_1 weighted image while echo image refers to the T_2 weighted image. Although the gradient echos and SSFP techniques generally suffer from

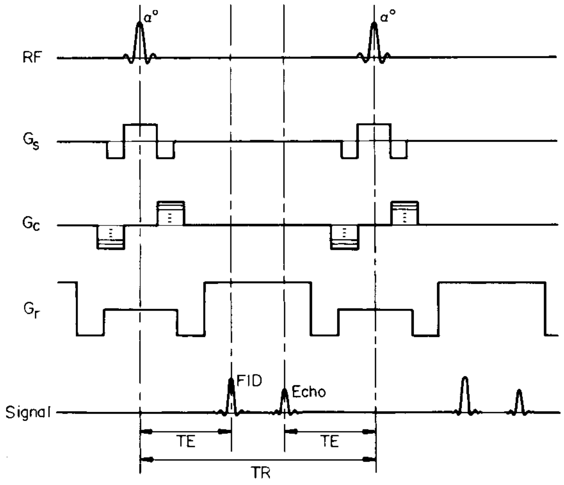


FIGURE 27 Imaging pulse sequence of the fast gradient echo imaging or its variation known as SSFP (steady-state free precession) imaging technique. With this pulse sequence, both the FID and the echo images, each of which has a characteristic contrast (namely, T_1 contrast in the FID image and strong T_2 contrast in the echo image) can be obtained simultaneously.

the susceptibility artifact, the methods nevertheless are becoming more widely used because of the advantages of speed and good image quality.

SEE ALSO THE FOLLOWING ARTICLES

Image-Guided Surgery • Image Processing • Magnetic Resonance in Medicine • Nondestructive Testing • Nuclear Magnetic Resonance (NMR) • Radionuclide Imaging Techniques, Clinical • X-Ray Analysis

BIBLIOGRAPHY

Barrett, H. H., and Swindell, W. (1981). "Radiological Imaging: The Theory of Image Formation, Detection, and Processing," Academic Press, New York.

Bushong, S. (2000). "Essentials of Medical Imaging: Computed Tomography," McGraw-Hill Professional, New York.

Carson, R. E. ( ). "Quantitative Functional Brain Imaging with Positron Emission Tomography," Academic Press, San Diego.

Cho, Z. H. (ed.) (1974). "Special Issue on Physical and Mathematical Aspects of 3-D Image Reconstruction," IEEE Trans. Nucl. Sci. NS-21, No. 2. Inst. Electr. Electron. Eng., New York.

Cho, Z. H. (ed.) (1976). "Special Issue on Advances in Picture Reconstruction Theory and Applications," Comput. Biol. Med. Vol. 6, No. 4. Pergamon Press, Oxford.

Cho, Z. H., and Nalcioglu, O. (eds.) (1984). "Special Issue on Physics and Engineering in Nuclear Magnetic Resonance Imaging," IEEE Trans. Nucl. Sci. NS-31, No. 4. Inst. Electr. Electron. Eng., New York.

Gardner, R. J. (1995). "Geometric Tomography," Cambridge Univ. Press, Cambridge, UK.

Herman, G. T. (1980). "Image Reconstruction from Projection," Academic Press, New York.

Herman, G., and Kuba, A. (1999). "Discrete Tomography: Foundations, Algorithms, and Applications," Birkhauser Boston, Cambridge, Massachusetts.

Kimmich, R. (1997). "NMR: Tomography, Diffusometry, Relaxometry," Springer-Verlag, Berlin/New York.

Macovski, A. (1983). "Medical Imaging Systems," Prentice-Hall, Englewood Cliffs, New Jersey.

Mansfield, P., and Morris, P. G. (1982). "NMR Imaging in Biomedicine," Academic Press, New York.

Nalcioglu, O., and Cho, Z. H. (eds.) (1984). "Selected Topics in Image Science," Springer-Verlag, Berlin/New York.

Newton, T. H., and Potts, D. G. (eds.) (1981). "Radiology of the Skull and Brain: Technical Aspects of Computed Tomography," Vol. 5. Mosby, St. Louis, Missouri.

Partain, C. L., James, A. E., Rollo, F. D., and Price, R. R. (1983). "Nuclear Magnetic Resonance (NMR) Imaging," Saunders, Philadelphia.

Yeon, K. M., Li, G.-Z., and Wol, C. (1995). "Computed Tomography: State of the Art and Future Applications," Springer-Verlag, Berlin/New York.

Ultrafast Spectroscopy and its Applications

M. Hayashi
Y. M. Chang
T. K. Wang
National Taiwan University

S. H. Lin
Academia Sinica, Taiwan

Joseph L. Knee
Wesleyan University

I. Ultrafast Spectroscopy Techniques
II. Gas and Liquid Phases
III. Biological Systems
IV. Semiconductors

GLOSSARY

Cage recombination The solvent molecules around a solute molecule act as a "cage" which restricts the motion of the solute in the photodissociation reaction.

Coherence in materials Due to the uncertainty principle, the energy distribution of photons in a pulse laser shows a finite bandwidth. Thus, a pulse laser can excite several quantum states of materials within the bandwidth at the same time. Such an excited state can be expressed in terms of a linear combination of these states and develops in time with a definite phase if there is no disturbance from the environments.

Dephasing The process in which the coherence of the oscillator decreases due to stochastic interaction with its surroundings.

Femtosecond Unit of time that is equal to 10^{-15} sec.

Laser-induced fluorescence Spectroscopic technique in which the sample system is excited by laser pulse and subsequently this excitation event is detected by emitted photons from the excited state.

Nanostructure The dimension of the material structure from one to one hundred nanometer (10^{-9} m).

Photodissociation The bond breaking process initiated by the absorption of optical photons.

Pump-probe spectroscopy Time-domain spectroscopic technique in which an intensive pump pulse is used to perturb some physical property of the sample system and subsequently a weak time-delayed probe pulse is applied to monitor the pump induced effect as a function of the delay time.

Solvation The combined effects of solvent molecules to the chemical reactions in solutions.

Ultrafast General term to describe both picosecond and femtosecond time domain.

Vibrational coupling The interaction between different

vibrational modes, which is responsible for vibrational energy relaxation and dephasing.

THE PROGRESS of both technology and applications in the field of ultrafast processes over the last 20 years has broadened the scope of ultrafast sciences. There are various ways to define the term "ultrafast." Here "ultrafast" means "femtosecond (fsec)," which is a unit of time equal to 10^{-15} seconds. This timescale becomes accessible due to progress in the generation, amplification, and measurement of ultrashort light pulses. With the advent of all-solid-state femtosecond laser sources and the extension of laser wavelengths via frequency conversion techniques, one can now obtain a variety of high-performance sources for extremely short light pulses.

Ultrafast laser pulse spectroscopy is a field of study wherein the unique properties of modern pulsed lasers are used to investigate the ultrafast dynamics of fundamental processes in physics, chemistry, and biology. The processes being investigated by femtosecond spectroscopic techniques can be categorized into three broad groups: (1) femtochemistry in gas and liquid phases; (2) electron-hole-phonon interaction in semiconductors and nanostructures; and (3) ultrafast electron and energy transfer reactions in biological systems. Figure 1 exhibits approximate timescales that are relevant to physical, chemical, and biological changes in ultrafast time region. Clearly, a number of important fundamental mechanisms in chemical, biological, and semiconductor systems can be investigated with ultrafast laser spectroscopy, because they take place in the femto- and picosecond timescales. One of the most extraordinary features in ultrafast spectroscopy is not only its ability to measure the rate constants of ultrafast processes, but also its ability to generate and observe coherence in various materials.

The study of ultrafast elementary dynamics is no longer limited to model systems but has now been extended to systems of great complexity that are of interest to condensed materials including biological systems. One can expect to witness "coherence" of such a wide range of materials. Several promising challenges to even control coherence have already been started utilizing controllable properties of a femtosecond pulse such as the central frequency, the width, and the phase together with an iterative algorithm method. These challenges will lead to extensions and new directions in femtosecond light-matter control based on the temporal coherence of light and its interference with matter waves.

I. ULTRAFAST SPECTROSCOPY TECHNIQUES

A. Pump-Probe Techniques

Development of ultrafast lasers has been an area of continued intense activity in last three decades. Several techniques have been developed to generate ultrashort pulses. So far the availability of ultrafast lasers with a wide range of pulse widths, wavelengths, pulse energy, and pulse repetition rate makes it possible not only to investigate a broad range of physical, chemical, and biological phenomena using ultrafast optical spectroscopy, but also to investigate practical applications of ultrafast technology.

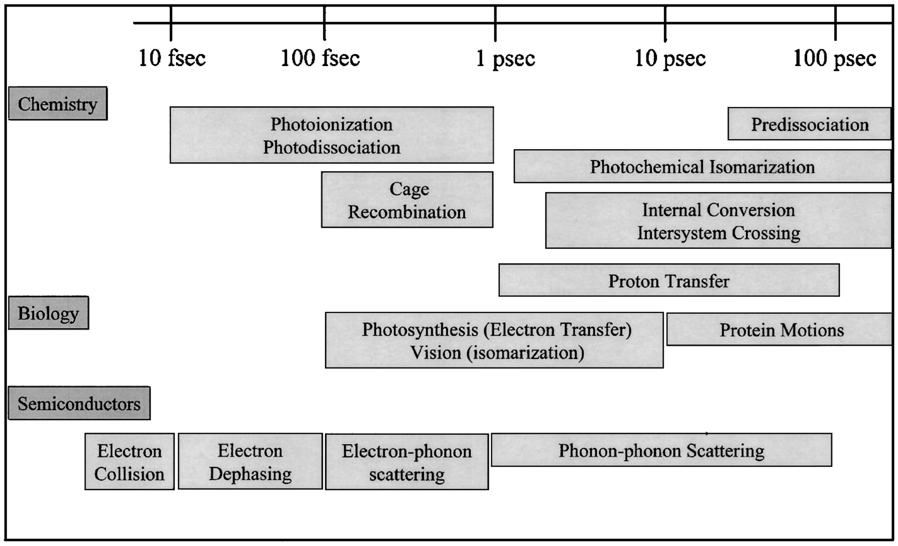


FIGURE 1 Typical timescales relevant to physical, chemical, and biological changes. The extension of the presented time range is not intended to imply upper or lower limits on some of the processes.

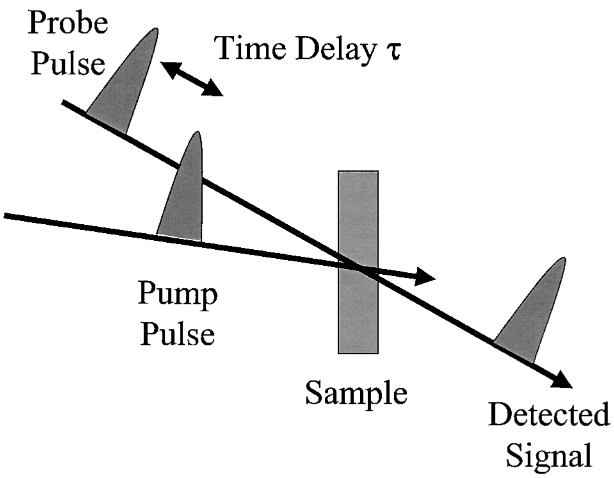


FIGURE 2 An illumination of typical pump-probe technique used in ultrafast spectroscopies. The delay time between the pump and probe pulses is controlled by varying the difference of path length of two laser pulses. The measured signal can be the linear or nonlinear reflection/transmission, Raman scattering, luminescence, or any other signal from the probe pulse.

Figure 2 illustrates the basic concept of a typical pump-probe spectroscopy used in most ultrafast spectroscopy techniques. In its simplest form the output pulse train of an ultrafast laser is divided in two by a beam splitter. One pulse in train (called pump) first excites the sample under investigation. The second pulse train (called probe) will probe the sample with a suitable time delay with respect to the pump by introducing an optical delay in its path and some optical property (e.g., reflectivity, absorption, Raman scattering, luminescence, optical nonlinear responses) of the sample is then detected to investigate the changes produced by the pump. In most of the time-resolved pump-probe experiments, the time resolution is limited only by the pulse width of the laser or the jitter between the laser systems.

B. Direct Time-Resolution Methods

The luminescence techniques have already become standard tools to investigate not only the ultrafast relaxation processes, but also the ultrafast transport processes in condensed matters. The interpretation of luminescence results may in some cases be simpler than that of pump-probe results because the latter is a nonlinear technique while luminescence is a linear technique. In other words, the observed results do not contain information of transient dynamics of the initial states of materials.

The best technique for direct luminescence measurement is the use of streak cameras, which can now provide a subpicosecond time resolution in single shot operation. In essence, it takes a snapshot in time of an event with such a resolution. It can record entire dynamics using only a single femtosecond laser pulse.

The pump-induced luminescence signal is collected and focused on a nonlinear crystal and the probe pulse is suitably delayed and also focused on the nonlinear crystal to overlap the luminescence spot. This technique provides a means for measuring low-level luminescence signal with time resolution limited by the laser pulse width, and has been applied to a wide variety of investigations in materials.

II. GAS AND LIQUID PHASES

A. Direct Observation of NaI Dissociation Reaction

Since most of the dissociation reaction of an isolated molecule occurs in a very short timescale (~100 fsec), the direct observation of its transition from reagents to products has to wait until the discovery of femtosecond laser technology. With femtosecond wave packet spectroscopy, the trajectories of particles can be monitored during their motions on a potential energy surface. Zewail's group in 1989, for the first time, directly followed the evolution in space and time of such trajectories during the breakage of a chemical bond in the dissociation of sodium iodide.

Figure 3 shows a schematic representation of the potential energy surface for the NaI reaction as well as the wave packet motion in its dissociation reaction. Two potential curves (covalent character, Na + I, and ionic character, $Na^+ + I^-$) intersect at an internuclear separation of 7Å, R_x. When NaI is promoted with a femtosecond (pump) pulse to the covalent curve at R = 2.7Å, the wave packet moves toward the intersection point, and is split into two parts. One part continues moving along the covalent potential curve. The other part is transformed into the ionic

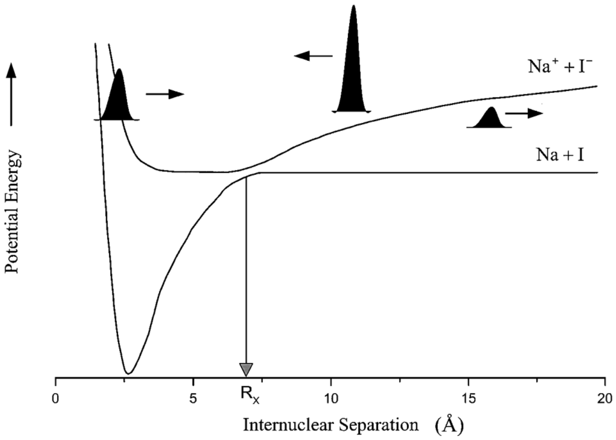


FIGURE 3 Schematic diagram of the wave packet motion in the dissociation reaction of NaI.

character and is eventually confined by the ionic potential curve to return with a period determined by the shape of the combined potential energy curve. Using wavelength tunable femtosecond probe pulses, the oscillation of this activated complex as well as the creation of Na + I can be monitored. In the experiments, the ~50 fsec pump pulse at 310 nm was generated by frequency doubling the amplified output of a colliding pulse mode-locked dye laser. The tunability on the probe pulse was achieved by generating a white-light continuum. The pump and probe beams were focused in an NaI reaction chamber and the laser-induced fluorescence (LIF) was recorded as a function of the time delay between the pump and probe pulse.

At the probe wavelength of 615 nm, the photon energy of the probe beam is off the absorption resonance of free Na atoms. The absorption of the probe takes from a unique position along the reaction coordinate to an upper excited state to emit fluorescence, resulting in probing this transition state. Each time the wave packet moves in the probing region of the potential energy surface, which is defined by the probe wavelength, the LIF signal displays a maximum. The transient signal thus exhibits an oscillatory feature, indicating that the wave packet moves back and forth along the combined potential energy surface. Furthermore, the peak value in the LIF signal decreases with the pump-probe time delay. It reflects the fact that each time the wave packet moves to the intersection point R_x, a portion of wave packet moves forward to produce products, leading to the decrease of the wave packet amplitude. At the probe wavelength of 590 nm, the probe photon is on the absorption resonance of free Na atoms. The transient LIF signal then reflects the production of free Na atoms. The LIF signal, on the other hand, exhibits a stepwise increase. This indicates the fact that the production of Na is synchronized with the oscillatory LIF signal at the probe wavelength of 615 nm. The combination of the oscillatory feature of the wave packet motion in the transition state and the stepwise increase in the products clearly shows the direct evidence of femtosecond dissociation reaction.

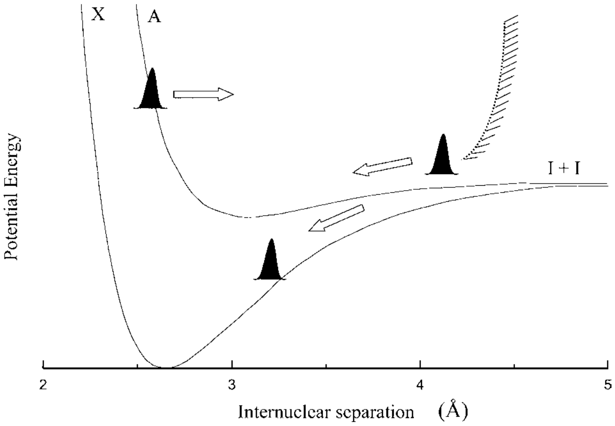


FIGURE 4 The potential energy surfaces (X and A) of the iodine molecule with snapshots of a wave packet motion on the A state. The dashed line in conjunction with the dashed area indicates the equivalent solvent barrier caused by the repulsion with Ar atoms.

B. Femtosecond Dynamics in $I_2 \cdot Ar_n$: Coherent Recombination and Dissociation

For chemical reactions in solution, the solvent exerts an important influence on the elementary processes of bond making and breaking. The solvent may, for example, enhance bond formation by trapping reactive species in a solvent cage on the reaction timescale, or it may act as a chaperone that stabilizes energetic species. Because of the complexity of solvent motions in liquid phases, reactive species embedded inside cluster cages become a model system to provide insights into such processes occurring in bulk condensed phases.

In $I_2 \cdot Ar_n$ clusters, neutral iodine molecules are encaged in argon clusters in a molecular beam. As shown in Fig. 4, dissociation was initiated by a femtosecond laser pulse which prepares a wave packet on the A state above its dissociation limit. On the A potential, the dissociation in the gas phase is direct and results in a large available translational energy for the I atoms. In a solvent cage, there is a solvent barrier to dissociation. In the pump-probe experiment of this system, iodine molecules were first excited at 614 nm to the A state. The probe-induced fluorescence then exhibits a peak at time zero. The following decay reflects the fact that the wave packet is moving into a region of large separations where iodine can no longer absorb the probe photons. In about 250 fsec, the signal drops to almost zero. A fast and prompt recovery was then observed in another ~300 fsec, which indicates that the wave packet coherently moves back to optical regions of the probe. This recovery is comparable to the coherent motion of isolated I_2 in a bound potential. Here it is the argon cage that provides the outer boundary of the potential well. At longer times, the signal starts to increase almost linearly and much more slowly. The picture described above has been confirmed by molecular dynamics simulations. These results demonstrate that in argon clusters, the motion of the wave packet is dramatically different from that of bare iodine and clearly reflects that the initial coherence and the caging dynamics.

C. Solvation Dynamics in Polar Liquids

The primary focus of the time-resolved studies in solvation has been its dynamics of ionic and polar solutes in strongly dipolar solvents, since it is the dynamics of such polar solvation that is expected to be important in determining

the kinetics of solution-phase charge transfer reactions. Subsequent to a perturbation of the charge distribution of a solute, solvent molecules reorganize so as to lower the free energy of solvation. The initial stage of solvation dynamics (<300 fsec) is the inertial motion of the solvent molecules and represents the relaxation of well over half of the solvation energy. The friction effects are manifested by the subsequent diffusive motion of solvent molecules. The direct experimental observation of these solvation processes is based on time resolving the fluorescence Stokes shift, that is, the shift in fluorescence frequency to longer wavelength as polar solvent molecules reorganize to accommodate the newly created excited-state charge distribution. The time-resolved fluorescence spectrum in femtosecond timescale is obtained by frequency up-conversion spectroscopy. In the experiments, the fluorescence induced by the pump pulse is collected and is mixing with the probe pulse in a nonlinear crystal. The sum-frequency spectrum produced by the fluorescence and probe photons is monitored as a function of the pump-probe time delay. With this technique, the femtosecond time resolution of the fluorescence spectrum can then be achieved.

The fluorescence Stokes shift function $S(t)$, for example a coumarin dye in water, obtained by experiment can then be obtained. Here, $S(t)$ is a normalized function describing the process toward equilibrium: $S(t) = [v(t) - v(\infty)]/[v(0) - v(\infty)]$, where $v(t)$ is a characteristic frequency of the fluorescence spectrum, for example, its first moment, at time t. The initial solvent dynamics is responsible more than 50% of relaxation and is completed less than 50 fsec. It results from free, inertial, small-amplitude motions of the solvent molecules within the potential wells that they initially occupy. Following the shape break in the response, the solvent begins to restructure itself, and energy flows irreversibly into the solvent and the behavior is now dissipative. This lack of dissipation during the initial solvation epoch is of significance for chemical processes in which coherent contributions can be important. Examples are electron or energy transfer processes mediated by bridge states.

D. Two-Dimensional Spectroscopy

Multidimensional nuclear magnetic resonance (NMR) techniques, since their inception in the 1970s, have turned into a powerful tool that provides structures and dynamics of very complex molecules, crystals, and proteins. By extending these ideas to the optical regime, it is possible to come up with novel classes of spectroscopies that can probe complex vibrational motions. Its realization, however, has to wait for the recent breakthrough in femtosecond laser techniques, which has provided necessary impulsive excitation to overcome tremendous line broadening caused by the ultrafast dephasing processes in the condensed phases.

In two-dimensional Raman spectroscopy, it is the fifth-order nonlinear optical response responsible for the observed signal. Between the preparation of the initial vibrational coherence and the probing step in traditional one-dimensional third-order Raman spectroscopy, there is a second interaction with an additional pair of light fields that transfers the initial vibrational coherence to a second vibrational coherence. The second coherence is then probed in the final step. Thus, it has two controllable time periods, and the signal $S(\tau_2, \tau_4)$ is governed by the fifth-order response function. The 2-D signal can be displayed in the frequency domain by defining the Fourier transform of $S(\tau_2, \tau_4)$ with respective to τ_2 and τ_4, i.e., $S(\omega_2, \omega_4)$. Like in multidimensional NMR spectroscopy, the cross peaks in the 2-D Raman spectrum should provide a direct probe for intermode couplings, which can be used as a sensitive probe to local structural information. Future theoretical investigation will be focused on connecting the detailed shape of this 2-D Raman spectrum to vibrational coupling dynamics.

III. BIOLOGICAL SYSTEMS

A. Ultrafast Biological Dynamics

Quite recently, important contributions have been made in the study of the elementary steps in vision, photosynthesis, protein dynamics, and electron and proton transport in DMA. An ultrafast event, bond cleavage, or electron transfer also occurs in proteins such as those of photosynthetic reaction centers and antennas, hemoglobins, cytochromes, and rhodopsin. One of the most remarkable achievements in the studies of ultrafast biological events has been due to the fact that the existence of global and coherent nuclear motions has been revealed in these complex systems. Quantum beats can be observed in these protein systems using femtosecond spectroscopy. This impact has changed the way of looking at the complexity of these systems and opened the door that leads to the bold challenge to understand what makes biological systems different from nonbiological systems.

B. Photosynthesis

1. Electron Transfer in Photosynthetic Reaction Centers

The study of the primary process of photosynthesis has been performed using transient absorption techniques. In this case, the reaction is monitored by measuring the absorption of the probe by the reactant, the product, or

perhaps an intermediate that exists only transiently as the reaction proceeds. The most thoroughly investigated of bacterial photosynthetic systems have been those bacterial photosynthetic systems such as the purple nonsulfur bacteria *Rhodopseudomonas* and *Rhodobacter*. A major breakthrough for the bacterial reaction center occurred about ten years ago when its structure was determined at high resolution. The bacterial reaction center of *Rhodobacter sphaeroides*, for example, is a pigment-protein complex designed to convert optical excitation into the initial electron transfer event in photosynthesis. The reaction center contains six chlorophyll-like pigments arranged with approximate C_2 symmetry. The relative arrangement of these pigments within the protein matrix is well-known as the special pair—two accessory bacteriochlorophylls and two bacteriopheophytins.

It has been shown that the initial process is an electron transfer to form a charge-separated special pair. This charge separation has been measured to occur less than 100 fsec and it has also been shown that subsequent electron transfer reactions from the charge-separated special pair to one of the two bacteriopheophytins take place within about 1–4 psec depending on temperature. The results of femtosecond experiments together with the bacterial reaction center structural determination raise the challenge of the relating mechanism of the primary electron transfer to the known structures. This fascinating problem is still under intensive investigation.

Coherent nuclear motions of bacterial photosynthetic reaction center have been visualized observing pump-probe-stimulated emission with 80-fsec pulse lasers.

Recent intensive investigations using pump-probe techniques and femtosecond anisotropy measurements have revealed that ultrafast energy transfer ($\sim$80 fsec) from one of the two accessory bacteriochlorophylls to the special pair and internal conversion ($\sim$120 fsec) within the special pair occur prior to electron transfer from the special pair to one of the two bacteriopheophytins.

2. Energy Transfer in Antenna

The role of the antenna pigments is to collect light energy from the sun and transfer it to reaction centers. In green plants and algae, these pigments are found in the thylakoid membranes of chloroplasts. In photosynthetic bacteria, the pigments are found in intracytoplasmic membranes or in special vesicles in the cell. The principle antenna pigments are chlorophyll *a* and *b* in plants, chlorophyll *c* in some algae, and bacteriochlorophyll *a*, *b*, or *c* in bacteria. Other pigments, called accessory pigments, are carotenoids and phycobiliproteins. Following light absorption by a pigment molecule, the electronic excitation is transferred until it is trapped by a reaction center.

Cyanobacteria and red algae utilize antenna pigments called phycobilins packed into complexes called phycobilisomes, which are attached to the photosynthetic membranes. Phycobilisomes contain several hundred billion chromophores, linear tetrapyrroles attached to the protein. These proteins are organized into disks that are themselves stacked into rods, with disks containing shorter wavelength pigments on one end, and longer wavelength pigments at the other end next to a central core. Thus the shorter wavelength absorbers: phycoerythrins (PE), 570 nm; are on the outside, phycocyanins (PC), 630 nm; within them, allophycocyanins (APC), 650 nm; in the core, followed by chlorophyll *a* inside the photosynthetic membrane.

The X-ray crystal structure of the trimeric aggregation state of APC isolated from the cyanobacterium *Spirulina platensis* is well-characterized: the APC trimer can be described as a C_3-symmetric, ring-like homotrimer of α and β polypeptide monomers. The phycocyanobilin chromophore (PCB) is arranged as dimers formed across the polypeptide interfaces between adjacent α, β polypeptide monomers.

The absorption spectra of the homogenous monomers for APC are almost identical to that of the α-subunit, and it has an absorption maximum at 614 nm while trimers have a sharp maximum at 650 nm and a prominent shoulder at about 610–620 nm. The origin of electronic states involved in steady-state spectroscopy is needed for understanding of femtosecond time-resolved spectra, especially when one relates the mechanism of the kinetics appearing in the time-resolved spectra to nonradiative transitions among the electronic states. If electronic Hamiltonian is fully diagonalized, in other words, delocalized electronic wavefunction is used to describe the system, nonradiative transitions between the same spin multiplicity should be due to the break-down of the adiabatic approximation, i.e., internal conversion (IC) process. In this case, to estimate the electronic coupling constant, the Förster or Marcus type of rate constant cannot be used. Vibrational properties are also needed to estimate the electronic coupling constant. Vibrational frequencies involved in optical transition of APC trimer can be obtained by using a pump-probe technique with 20-fsec laser pulses. Figure 5 shows femtosecond pump-probe signals of APC trimer at room temperature as a function of the probing wavenumber. Each signal clearly shows oscillatory features called "quantum beats" accompanied by with rising or decaying components. Insets of panel A exhibit oscillatory components extracted from the observed signals. Performing Fourier transform analysis to these oscillatory components provides possible vibrational frequencies involved in the optical transition. This information can be used to estimate the electronic coupling constant.

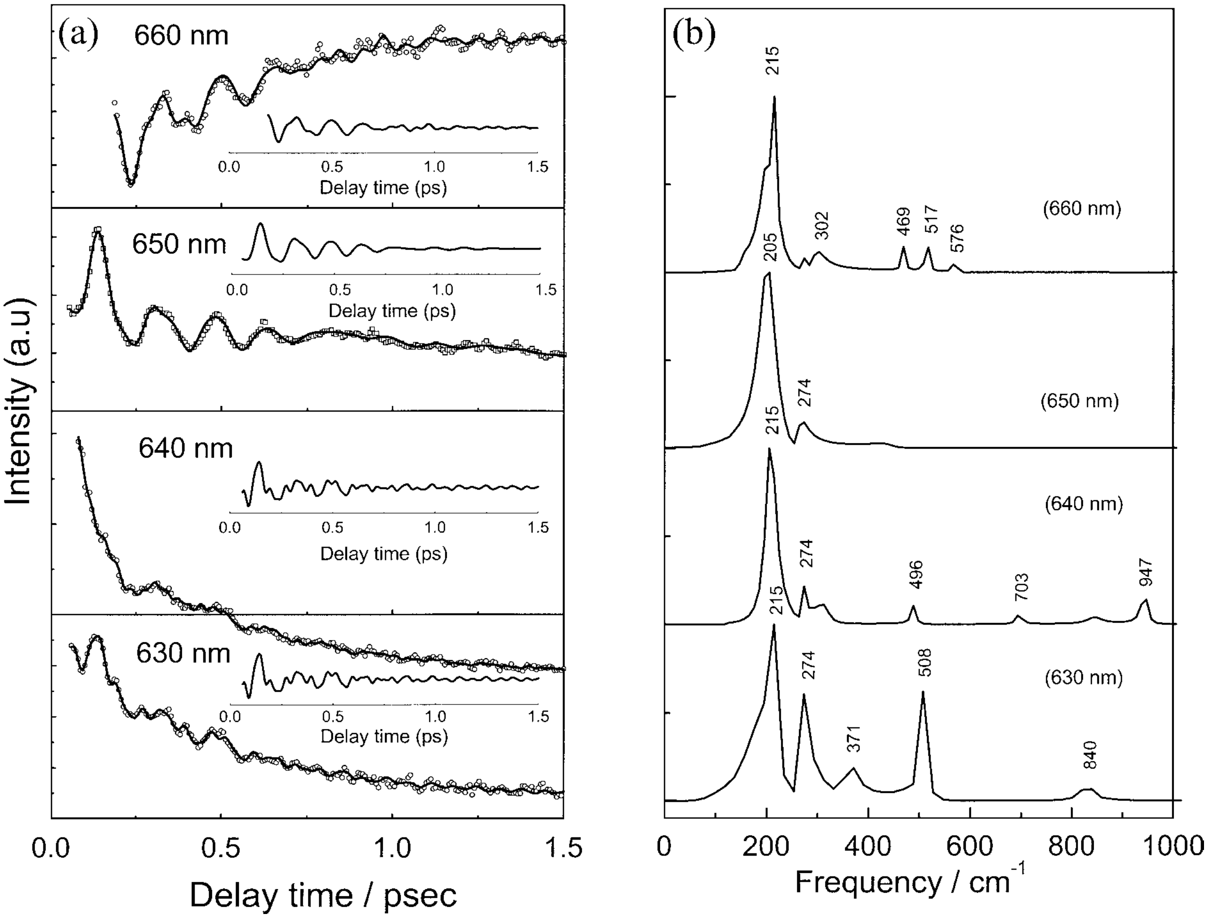


FIGURE 5 (a) Quantum beats observed in pump-probe signals of APC trimer at room temperature. Signals were obtained as a function of delay time between the pump and probe pulses as well as the probe wavelength. Insets are oscillatory components extracted from the observed signals. (b) Fourier transform analysis of the oscillatory components.

Femtosecond anisotropy measurements can provide useful information on the origin of rising and decay components appearing in the observed signals. Anisotropy can be obtained by monitoring the spectrally resolved probe transmission as a function of temporal delay and polarization relative to the pump. The anisotropy can be calculated from the pump-probe signals as

$$r(\tau) = \frac{S_{\parallel}(\tau) - S_{\perp}(\tau)}{S_{\parallel}(\tau) + 2S_{\perp}(\tau)},$$

where $S_{\parallel}(\tau)$ and $S_{\perp}(\tau)$ are pump-probe transients obtained with parallel and perpendicular pump and probe polarizations, respectively. Figure 6 shows obtained femtosecond anisotropy signals of APC trimer excited at 620 nm and probed at 660 and 620 nm. The obtained anisotropy indicates that there exist at least two electronic states between 660 and 620 nm. Fitting the anisotropy decay signal of probing at 660 nm to a single exponential decay function yields approximately 350 fsec. This result implies that the electronically excited state via the 620 nm pump-pulse undergoes nonradiative transition to one of the lower electronic states within 350 fsec. The origin and nature of the electronic states of the APC trimer are still under extensive investigation.

IV. SEMICONDUCTORS

A. Ultrafast Dynamics

The scattering and relaxation processes of photoexcited carriers in semiconductors and their nanostructures, particularly the III–V group semiconductor and its quantum well structures, have been extensively investigated using the ultrashort spectroscopy techniques described above. Intrinsic semiconductors are characterized by their band structure, which contains a bandgap separating the occupied valence band and the empty conduction band. The

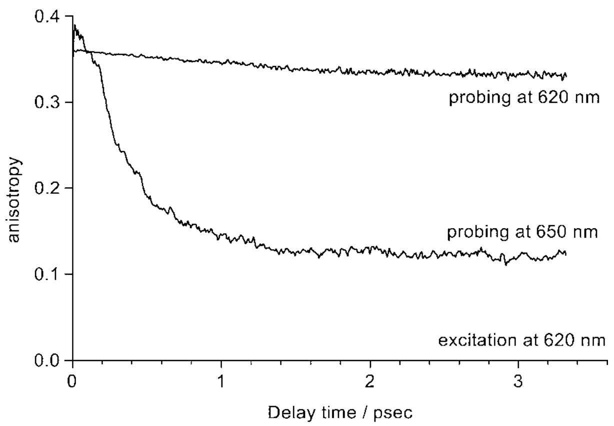


FIGURE 6 Anisotropy calculated from pump-probe signals of APC trimer.

band structure governs many of the electric and optical properties of semiconductors. It is an essential concept to lead people to understand the scattering and relaxation processes of carriers in semiconductors. For example, bulk GaAs has a direct bandgap located at the center of the Brillouin zone. The conduction band has subsidiary minima at some symmetry points in the Brillouin zone. The subsidiary valleys play an important role in the transport or scattering of photoexcited carriers in GaAs. Near the Γ valley, the carrier energy dispersion with its wavevector E (***k***) is a parabolic relationship. The electron states in this valley are twofold degenerate and are primarily s-like near the minimum, with an increasing admixture of p-character at higher energies. On the other hand, the valence band in GaAs is fourfold degenerate at the zone center and is split into the Heavy-Hole and the Light-Hole bands at larger ***k*** values. This degeneracy can be further split by the surface or interface stress or quantum confinement in semiconductor nanostructures. This complicated band structure is necessary and essential to understand many of the ultrafast phenomena regarding the relaxation of photoexcited carriers of semiconductors in next sections.

Scientists now can easily sandwich a semiconductor with a smaller bandgap (GaAs) between layers of a larger band-gap semiconductor (AlGaAs) with epitaxial growth techniques such as molecular beam epitaxy (MBE). One can even manipulate the thickness of the smaller band-gap material to be comparable to or smaller than the carrier de Broglie wavelength. In this case, quantum phenomena cannot be ignored anymore. A multiple quantum well (MQW) structure consisting of several such periods can also be made and treated as an independent well if the barrier height and thickness are sufficiently large. This kind of semiconductor nanostructure leads to a profound change in the properties of the semiconductor due to the quantum confinement effects. The changes result obviously from the fact that carriers are confined along the growth direction of the structure, but are free to move in lateral directions. The dimensionality of the semiconductor nanostructures can be further reduced. The one and zero dimensional systems, so-called quantum wire and quantum dot, have attracted a lot of attention recently and the area of this research can be expected to broaden and yield fascinating results that cannot be presently anticipated.

B. Scattering Processes in Semiconductors and Their Nanostructures

Whenever a semiconductor in thermodynamic equilibrium is excited by an ultrashort laser pulse, it undergoes several stages of energy and momentum relaxation before it returns once again to the thermodynamic equilibrium. The dynamics of photoexcited carriers (electrons and holes) are influenced by the interaction among them, as well as with the phonons, impurities, defects, and interfaces of the system. In the temporal region, one can classify those processes into four regimes: (1) coherent regime (<200 fsec); (2) non-thermal regime (<2 psec); (3) hot carrier regime (∼1–100 psec); and (4) isothermal regime (>100 psec). Figure 7 illustrates some of the involved ultrafast scattering processes in the dynamics of photoexcited carriers in semiconductors and their nanostructures.

Carrier-carrier scattering determines the exchange of energy and momentum between carriers. This scattering process is primarily responsible for the mail thermalization channel of photoexcited nonthermal carriers in the early stage. This process is mainly due to the Coulomb interaction between the electron-electron, hole-hole, and electron-hole. The difference in the masses between electrons and holes reduces the energy exchange between these two species. It was found that photoexcited

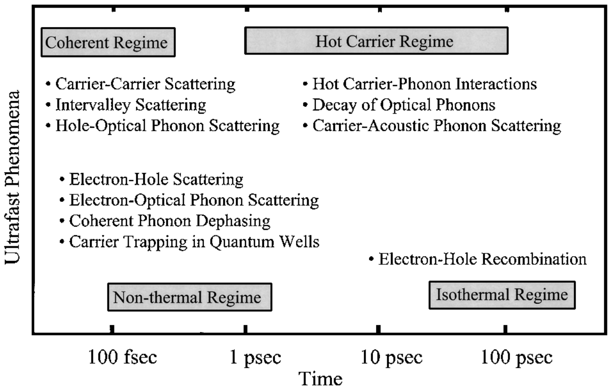


FIGURE 7 Typical scattering and energy relaxation processes occurred in photoexcited semiconductors and their nanostructures. Four main relaxation regimes are indicated in the time axis.

holes reach thermal equilibrium faster than the electrons. The long-range Coulomb interaction diverges in the absence of screening. Screening is generally treated as the plasmon-phonon coupling and occurs in the nonthermal regime.

Interaction of carriers with phonons plays a major role in the exchange of energy and momentum between carriers and the lattice, and hence determines the relaxation of photoexcited carriers and their transport properties in semiconductors. Polar optical phonons, particularly, play an important role in carrier-phonon scattering processes in group III–V semiconductors. The interaction between polar optical phonons and low energy carriers is described by the *Fröhlich interaction*, in which the polar coupling strength varies inversely with the wavevector $\boldsymbol{q}$ of the phonon being emitted or absorbed. Since the *Fröhlich interaction* rate is expected to vary $m^{1/2}$, the holes might be expected to interact more strongly with the LO phonons than the electrons, as indicated in Fig. 7. Note that the high-energy electrons and holes in subsidiary valleys can also interact with optical phonons through nonpolar optical deformation potential scattering. This interaction is also responsible for the intervalley scattering of electrons in the nonthermal regime. Interaction with both LA and TA phonons can also be important in the thermalization process of hot carriers. However, if the carrier has sufficient energy to emit an optical phonon, then the optical phonon scattering rate is generally considerably higher than for the acoustic phonons. In brief summary, carrier-carrier scattering is primarily responsible for redistributing the energy and leads to a thermalized distribution function of the carrier system. The characterized temperature of the carrier system can be higher than the lattice temperature, and may be different for electron and hole systems. Typically, the electrons and holes thermalize among themselves in hundreds of femtoseconds, while the electrons and holes achieve a common temperature in a couple of picoseconds. The carrier system finally reaches lattice temperatures in hundreds of picoseconds through the interaction with various phonons in the semiconductor.

In the isothermal regime, all the carriers and phonons are in thermal equilibrium with each other, i.e., they can be described by the same lattice temperature. However, there is still an excess of electrons and holes compared to the thermodynamic equilibrium. These excess electrons and holes must return to the thermodynamic equilibrium through either radiative recombination or nonradiative recombination.

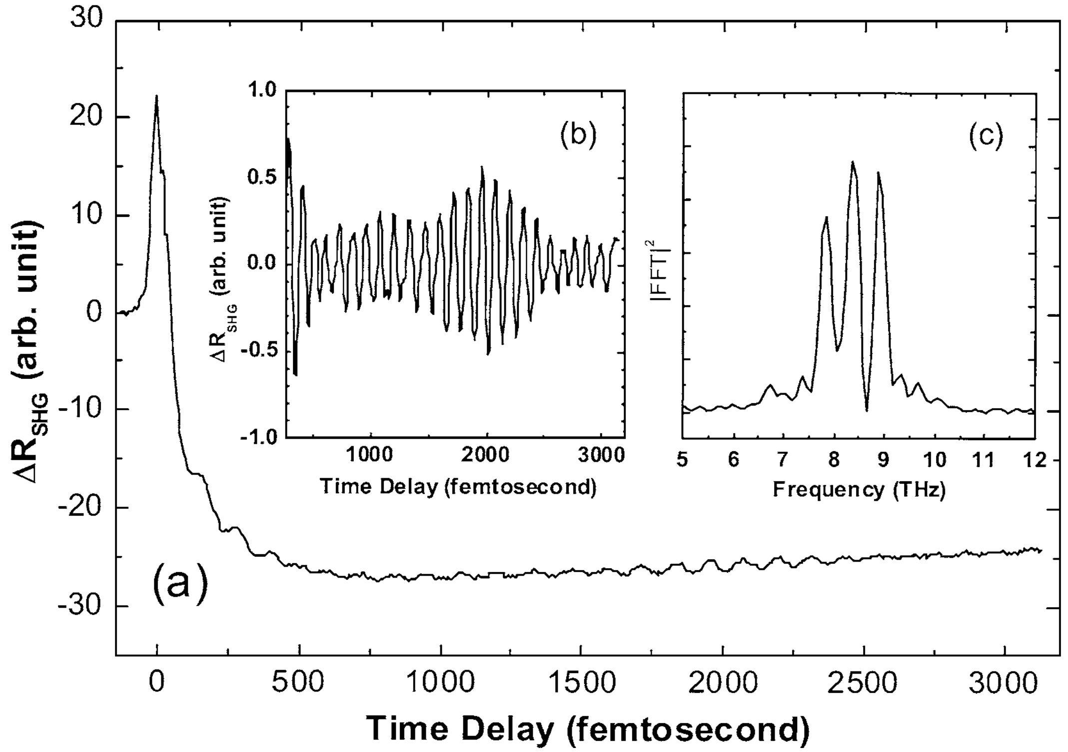


FIGURE 8 (a) Time-resolved second-harmonic generation spectroscopy reveals the generation of coherent phonon modes in native oxide covered GaAs (100) crystal. (b) Left inset shows the oscillatory part of the time domain data. (c) Right inset shows the Fourier power spectrum of the oscillatory part.

C. Dephasing of Coherent Phonons

In recent years, generation of coherent optical phonons in semiconductors, semimetals, and superconductors by femtosecond excitation has also received considerable attention. The coherent phonon oscillations are investigated with time-resolved linear reflectivity/transmission or nonlinear optics techniques. Figure 8 shows an example of the generation and detection of coherent phonons in GaAs (100) crystal. In that experiment, time-resolved second harmonic generation technique is used to probe the coherent phonon modes. Since the reflected second-harmonic generation has intrinsic surface and interface sensitivity, the measured time-resolved spectrum and its Fourier transform indicate that three phonon modes are impulsively generated by a femtosecond laser pulse (<50 fsec). The three phonon modes are assigned to the bulk LO phonon (8.8 THz), the native-oxide-GaAs interfacial phonon (8.4 THz), and the plasma-bulk LO phonon coupling mode (7.8 THz). The near surface depletion field screened by photoexcited carriers is responsible for an ultrafast longitudinal depolarization of the crystal lattice, which launches these phonon modes in the near surface region of GaAs. This ultrafast spectroscopy technique therefore allows one to investigate not only the dephasing mechanisms of phonons but also the influence of carrier density on the phonon modes. It clearly provides a powerful means of investigating phonons and their interactions with carriers in semiconductors directly in the time domain.

SEE ALSO THE FOLLOWING ARTICLES

Electron Transfer Reactions • Coherent Control of Chemical Reactions • Lasers, Ultrafast Pulse Technology • Luminescence • Molecular Beam Epitaxy, Semiconductors • Nanostructured Materials, Chemistry of • Nonlinear Optical Processes

BIBLIOGRAPHY

Fleming, G. R., Joo, T., and Cho, M. (1997). Femtosecond chemical dynamics in condensed phases. *Adv. Chem. Phys.* **101,** 141–180.

Hochstrasser, R. M., and Johnson, C. K. (1992). *In* "Ultrashort Laser Pulse: Generation and Applications" (W. Kaiser, ed.), 2nd edition, Springer-Verlag, Berlin and New York.

Martin, J. L., and Vos, M. H. (1992). Femtosecond biology. *Annu. Rev. Biophys. Biomol. Struct.* **21,** 199–222.

Mukamel, S., Piryatinski, A., and Chernyak, V. (1999). Two-dimensional raman echoes: femtosecond view of molecular structure and vibrational coherence. *Acc. Chem. Res.* **32,** 145–154.

Shah, J., ed. (1992). "Hot Carriers in Semiconductor Nanostructures: Physics and Applications," Academic Press, Boston.

Shah, J. (1998). "Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures," 2nd edition, Springer-Verlag, Berlin and New York.

Shank, C. V., Zakharchenya, B. P. (1992). "Spectroscopy of Nonequilibrium Electrons and Phonons," Elsevier, Amsterdam.

Zewail, A. H. (2000). Femtochemistry: atomic-scale dynamics of the chemical bond. *J. Phys. Chem. A* **104,** 5660–5694.

Zewail, A. H. (1994). "Femtochemistry: Ultrafast Dynamics of the Chemical Bond," World Scientific, Singapore.

X-Ray Analysis

Ron Jenkins
International Centre for Diffraction Rate

I. Use of X Rays for Materials Characterization
II. Origin and Properties of X Rays
III. X-Ray Wavelengths and Atomic Number
IV. X-Ray Fluorescence Spectrometry
V. Application of X-Ray Spectrometric Methods
VI. Trace Analysis by X-Ray Fluorescence Spectrometry
VII. The Crystalline State
VIII. X-Ray Powder Diffractometry
IX. Application of X-Ray Diffraction Methods
X. Combined Diffraction and Spectroscopic Techniques

GLOSSARY

Auger effect Rearrangement of atomic electrons, leading to the emission of further photoelectrons, that follows the ejection of a primary atomic electron.

Collimator Series of flat, high X-ray absorbing metal plates, used to limit the vertical divergence of a beam of X rays.

Crystal lattice Unique arrangement of atoms in a specific element or compound determined by the number, size, and bonding configuration of the atoms.

Crystallite orientation Tendency for individual crystallites in a powder sample to lay in a preferred direction because of their external shapes.

Energy-dispersive X-ray spectrometry Method for elemental analysis based on the separation of X-ray according to their energies, by use of a lithium-drifted silicon proportional detector.

JCPDS Powder Date File Standard library of approximately 46,000 single-phase diffraction patterns, used in qualitative powder diffractometry.

Mass absorption coefficient (or mass attenuation coefficient) Quantity with the units of square centimeters per gram, representing the absorption of a specific combination of elements for one fixed wavelength.

Photoelectric absorption Loss of energy of a primary X-ray photon following interaction with one or more atomic orbital electrons.

Wavelength-dispersive X-ray spectrometry Method for qualitative and quantitative elemental analysis using the wavelengths of characteristic X-ray emission

lines, in which a single diffracting crystal is used to separate the wavelengths from the sample.

X-Ray powder diffraction Technique for phase identification based on measurement of structure sensitive diffracted X-ray lines.

X-RAY ANALYSIS is the identification and characterization of materials through the use of X rays and includes the two techniques of spectrometry and diffractometry. X-Ray fluorescence spectrometry is a means of qualitatively and quantitatively determining elements by measurement of the wavelengths and intensities of characteristic emission wavelengths. The technique is applicable to all but the elements of very low atomic number, with sensitivities down to the low part per million level. X-Ray powder diffractometry uses diffraction data obtained by study of structure-dependent scattering phenomena. The shapes and positions of diffraction peaks can be used to establish a wide range of physical properties, including stress, texture, crystallinity, and particle size, but the main application is for phase identification. A large data base of pure single-phase materials is available so that rapid empirical file-searching techniques can be employed for measuring phases (compounds).

I. USE OF X RAYS FOR MATERIALS CHARACTERIZATION

X rays are a short-wavelength form of electromagnetic radiation discovered by Wilhelm Roentgen in 1895. X-Ray-based techniques provided important tools for the theoretical physicist in the first half of this century, and since the early 1950s they have found an increasing use in the fields of materials characterization. Today, the analytical techniques based on X-ray diffraction and X-ray spectrometry, both of which were first conceived almost 70 years ago, play a vital role in the analysis and study of inorganic and organic solids.

X-Ray photons are produced following the ejection of an inner orbital electron from an irradiated atom and subsequent transition of atomic orbital electrons from states of high to low energy. When a monochromatic beam of X-ray photons falls onto a given specimen, three basic phenomena may result: absorption, scatter, or fluorescence. The coherently scattered photons may undergo subsequent interference, leading in turn to the generation of diffraction maxima. The angles at which the diffraction maxima occur can be related to the spacings between planes of atoms in the crystal lattice, and hence X-ray-generated diffraction patterns can be used to study the structure of solid materials. Following the discovery of the diffraction of X-rays by Max von Laue in 1913, the use of this method for materials analysis has become very important both in industry and research, to the extent that, today, it is one of the most useful techniques available for the study of structure-dependent properties of materials.

X-Ray fluorescence spectrometry also has its beginnings back in the early part of the 20th century with the work of Moseley and others. The technique uses either the diffracting power of a single crystal to isolate narrow wavelength bands, or a proportional detector to isolate narrow energy bands, from the polychromatic beam characteristic radiation excited in the sample. The first of these methods is called wavelength-dispersive spectrometry, and the second, energy-dispersive spectrometry. Because of the known relationship between emission wavelength and atomic number, isolation of individual characteristic lines allows the unique identification of an element to be made, and elemental concentrations can be estimated from characteristic line intensities. Thus this technique is a means of materials characterization in terms of chemical composition.

X-Ray powder diffractometry involves characterization of materials by use of data that are dependent on the atomic arrangement in the crystal lattice. The technique uses single or multiphase (i.e., multicomponent) specimens comprising a random orientation of small crystallites, each of the order of 1–50 μm in diameter. Each crystallite in turn is made up of a regular, ordered array of atoms. An ordered arrangement of atoms (the crystal lattice) contains planes of high atomic density, which in turn means planes of high electron density. A monochromatic beam of X-ray photons will be scattered by these atomic electrons, and if the scattered photons interfere with each other, diffraction maxima may occur. In general, one diffracted line will occur for each unique set of planes in the lattice. A diffraction pattern is typically in the form of a graph of diffraction angle (or interplanar spacing) versus diffracted line intensity. The pattern is thus made up of a series of superimposed diffractograms, one for each unique phase in the specimen. Each of these unique patterns can act as an empirical "fingerprint" for the identification of the various phases, using pattern-recognition techniques based on a file of standard single-phase patterns. Quantitative phase analysis is also possible, albeit with some difficulty because of various experimental and other problems, not the least of which is of the large number of diffraction lines occurring from multiphase materials.

Although the major thrust of this review is to cover X-ray diffraction and X-ray spectroscopic techniques, these are by no means the only X-ray-based methods that are used for materials analysis and characterization. In addition to the many industrial and medical applications of

diagnostic X-ray absorption methods, X rays are also used in areas such as structure determination based on single-crystal techniques, space exploration and research, in lithography for the production of microelectronic circuits, and so on. One of the major limitations with the further application of X-ray methods is the inability to "focus" X-rays, as can be done with visible light rays. Although it is possible to partially reflect X-rays at low glancing angles, or to diffract an X-ray beam with a single crystal, these methods cause significant intensity loss and fall far short of providing the high-intensity, monochromatic beam that would be ideal for, for example, an X-ray microscope. Use of synchrotron radiation offers the potential of an intense, highly focused, coherent X-ray beam but has practical limitations due to size and cost.

The X-ray laser could, in principle, provide an attractive alternative, and since the discovery of the laser in 1960, the possibilities of such an X-ray laser have been discussed. Although major research efforts have been and are still being made to produce laser action in the far ultraviolet and soft X-ray regions, production of conditions to stimulate laser action in the X-ray region with a net positive gain is difficult, mainly because of the rapid decay rates and high-absorption cross sections that are experienced in practice. Many different experiments have been designed to circumvent these problems, and one of the more promising new approaches in soft X-ray laser development is based on rapid plasma cooling by radiation losses after the laser pulse. This leads to a fast recombination and collisional cascade into upper excited levels, while the lower excited levels are rapidly depopulated by radiative transitions. This causes a population inversion leading to an overall gain. A CO_2 laser is focused onto a target of Teffon and the resulting plasma confined to a field of 50–90 kG.

II. ORIGIN AND PROPERTIES OF X RAYS

If a high-energy particle, such as an electron, strikes a bound atomic electron, and the energy E of the particle is greater than the binding energy O of the atomic electron, it is possible that the atomic electron will absorb all of the energy from the particle. The primary particle disappears in the process, and its energy is transferred to the atomic electron, which is then ejected from its atomic position and leaves the atom with a kinetic energy E_{pe} equivalent to the difference between that of the initial particle and the binding energy of the atomic electron. The ejected electron is called a *photoelectron* and the interaction is referred to as the *photoelectric effect*. Figure 1 illustrates the process, and here an interaction with an electron in the K shell is indicated.

As long as the vacancy in the K shell exists, the atom is in an unstable state and there are two processes by which the atom can revert back to its original state. The first of these involves a rearrangement of the atomic electrons, as illustrated in Fig. 1d. This rearrangement does not involve the emission of X-ray wavelengths but does result in the emission of other photoelectrons from the atom. The effect is known as the Auger effect, and the emitted photoelectrons are called Auger electrons. These Auger electrons have little further consequence in this specific review on X-ray fluorescence and diffraction methods, but they do provide the basis of another important spectroscopic method called electron spectroscopy, which is especially important for surface and bonding studies.

The second process by which the excited atom can regain stability is by transference of an electron from one of the outer orbitals to fill the vacancy. This is illustrated in Fig. 1c. The energy difference between the initial and final states of the transferred electron may be given off in the form of an X-ray photon. The wavelength of the X-ray photon is inversely related to the energy of the photon. There are several different combinations of quantum numbers held by the electron in the initial and final states; hence several different X-ray wavelengths will be emitted from a given atom. For those vacancies giving rise to characteristic X-ray photons, a series of very simple selection rules can be used to define which electrons can be transferred: The principal quantum number n must change by 1, the angular quantum number l must change by 1, and the vector sum of $l + s$ must be a positive number changing by 1 or 0. In effect, this means that for the K series only $p \rightarrow s$ transitions are allowed, yielding two lines for each principal level change. Vacancies in the L level follow similar rules and give rise to L series lines. There are more of the L lines since $p \rightarrow s$, $s \rightarrow p$, and $d \rightarrow p$ transitions are all allowed within the selection rules.

In general, electron transitions to the K shell give between two and six K lines, and transitions to the L shell give about 12 strong to moderately strong L lines. As an example, Fig. 2 shows the L emission spectrum of gold ($Z = 79$). Since all emitted X-ray photons have energies proportional to the differences in the energy states of atomic electrons, the lines from a given element will be characteristic of that element. As will be shown in the next section, the ability to use a measured wavelength as a unique property of an excited element forms the basis of X-ray emission spectrometry.

All X-ray photons with wavelengths in the range generally employed for analytical work are absorbed to a lesser or greater extent by solid materials. If a monochromatic beam of X-ray photons of wavelength λ and intensity I_0 falls onto a material of thickness x and density ρ, the fraction of the beam transmitted I is given by

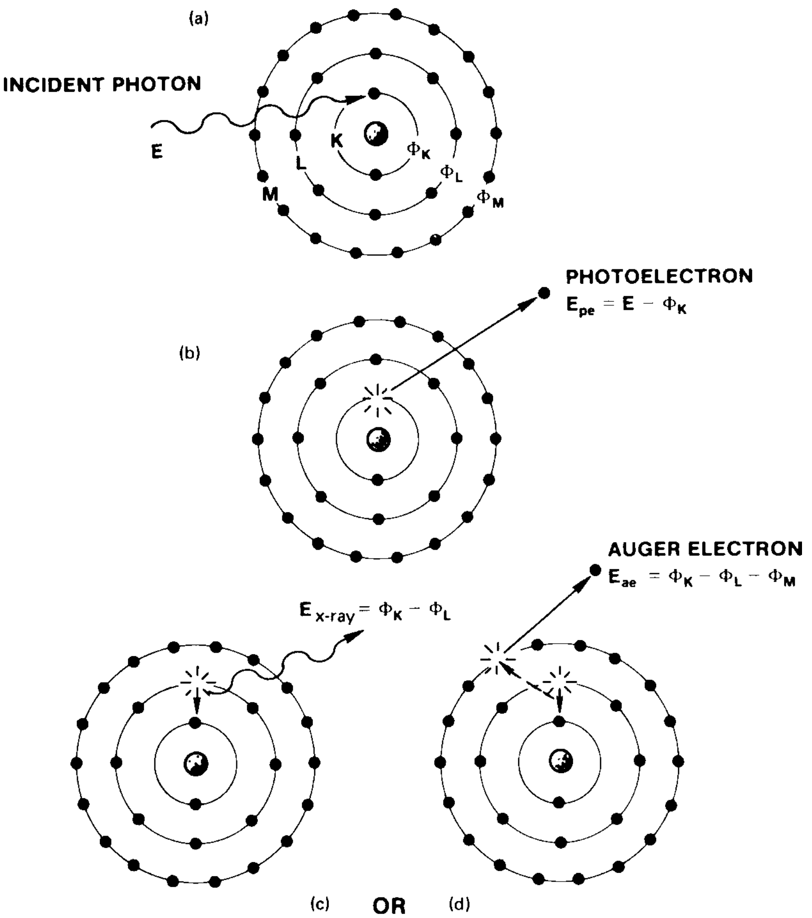


FIGURE 1 Production of characteristic X rays. (a) Excitation of a K-orbital electron producing a K vacancy. (b) Emission of a photoelectron producing a K vacancy. Deexcitation leads to (c) the emission of a characteristic K photon, or (d) an Auger electron.

$$I(\lambda) = I_0 \exp[-(\mu\rho x)] \quad (1)$$

where μ is the mass absorption coefficient of the absorber for the wavelength λ.

From Eq. (1) it will be apparent that a number of photons equal to $(I_0 - I)$ has been lost in the absorption process. Although a significant fraction of this loss may be due to scatter, by far the greater loss is due to the photoelectric effect. Photoelectric absorption will occur at each of the energy levels of the atom; thus the total photoelectric absorption will be determined by the sum of each individual absorption within a specific shell.

Where the absorber is made up of a number of different elements, as is usually the case, the total absorption is made up of the sum of the products of the individual elemental mass absorption coefficients and the weight fractions of the respective elements. This product is referred to as the total matrix absorption.

The value of the mass absorption referred to in Eq. (1) is a function of both the photoelectric absorption and the scatter. However, the photoelectric absorption influence is usually large in comparison with the scatter, and to all intents and purposes the mass absorption coefficient is equivalent to the photoelectric absorption. Because the photoelectric absorption is made up of absorption in the various atomic levels, it is an atomic-number-dependent function. A plot of the mass absorption coefficient as a function of wavelength contains a number of discontinuities called absorption edges. These occur at wavelengths corresponding to the binding energies of the electrons in the various subshells. The absorption discontinuities are a major source of nonlinearity between X-ray intensity and composition in both X-ray fluorescence and X-ray powder diffractometry.

Scattering will occur when an X-ray photon collides with the electrons of the absorbing element. Where this collision is elastic—that is, no energy is lost in the collision process—the scatter is said to be coherent (Rayleigh) scatter. Since no energy change is involved, the coherently scattered radiation will retain exactly the same wavelength

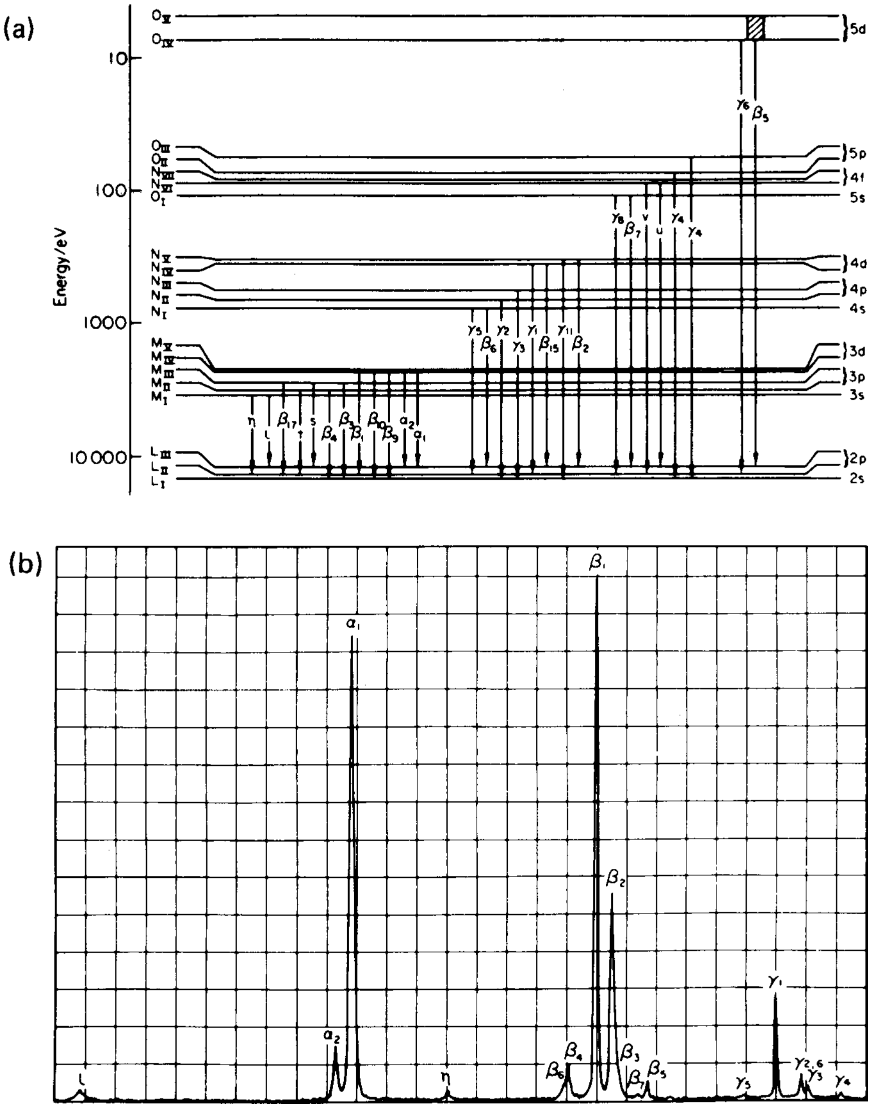


FIGURE 2 The L-series X rays from gold. (a) The energy-level diagram of gold shows all possible diagram lines and (b) the actual emission spectrum.

as that of the incident beam. It can also happen that the scattered photon gives up a small part of its energy during the collision, especially where the electron with which the photon collides is only loosely bound. In this instance the scatter is said to be incoherent (Compton scatter), and the wavelength of the incoherently scattered photon will be greater than that of the incident wavelength.

Under certain geometric conditions, wavelengths that are exactly in phase add to one another, or if exactly out of phase, cancel each other out. Under such conditions, coherently scattered photons may constructively interfere with each other, giving diffraction maxima. Since a crystal lattice consists of a regular arrangement of atoms forming layers of high atomic density, when a monochromatic beam of radiation falls onto these atomic layers scattering will occur. In order to satisfy the requirement for interference, it is necessary that the scattered waves originating from the individual atoms, that is, the scattering points, be in phase with one another. The geometric conditions for this condition to occur are illustrated in Fig. 3. Two parallel rays strike a set of crystal planes at an angle θ and are scattered as previously described. Reinforcement will occur when the difference in the path lengths of the two interfering waves is equal to a whole number of wavelengths. This path-length difference is equal to CB + BD, and since CB = BD = x, $n\lambda$ must equal $2x$ for reinforcement to occur, where n is an integer. It will also be seen that $x = d \sin\theta$, where d is the interplanar

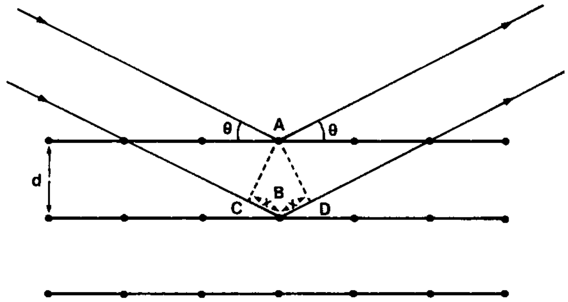


FIGURE 3 Condition for diffraction of X rays. A parallel beam of radiation falls onto planes in the crystal lattice separated by interplanar spacing *d*. If the path length difference of successive rays $2x$ is equal to an integral number of wavelengths, diffraction will occur at an angle θ.

spacing. Hence the overall condition for reinforcement is that

$$n\lambda = 2d \sin\theta \tag{2}$$

which is a statement of Bragg's law.

A powder diffractometer is essentially a device that presents a prepared specimen to a monochromatic beam of X rays and measures the angles of the diffracted lines. The actual instrument allows an angular range to be scanned, by rotation of the photon detector at twice the angular rotation speed of the specimen, thus maintaining the required geometrical condition. The specimen consists of a random distribution of crystallites, so that the appropriate planes will be in the correct orientation to diffract the incident wavelength each time the Bragg condition is satisfied. Thus, in X-ray diffraction, each peak angle value corresponds to a certain *d* spacing.

Use of the diffraction effect is made use of in the wavelength-dispersive spectrometer. In this instrument, one of a number of selectable single crystals, each of known *d* spacing, is used to disperse the polychromatic beam of characteristic wavelengths coming from the sample, such that each wavelength will diffract at a discrete angle. Different diffracting crystals are used, depending on the wavelength range sought and the required angular dispersion of the diffracted wavelengths. Thus, referring to Eq. (2), X-ray diffraction uses a fixed wavelength to estimate *d* spacings by measurement of diffraction angles, whereas X-ray spectrometry uses a fixed *d* spacing as a means of estimating the wavelength following measurement of the diffraction angles.

III. X-RAY WAVELENGTHS AND ATOMIC NUMBER

The wavelength and energy of a given X-ray photon are inversely proportional to each other. As indicated in Section II, the basis of the X-ray fluorescence technique lies in the relationship between the wavelength λ (or energy E) of the X-ray photons emitted by the sample element and atomic number Z, as given in Moseley's law:

$$1/\lambda = K(Z - s)^2 \tag{3}$$

In this expression K and s are constants that depend on the spectral series of the emission line in question. It has also been stated that not all vacancies result in the production of characteristic X-ray photons because of the competing Auger process involving internal rearrangement. The ratio of the number of vacancies resulting in the production of characteristic X-ray photons to the total number of vacancies created in the excitation process is called the fluorescent yield. The fluorescent yield is an important fundamental parameter, since it is one of the factors that will determine the absolute number of counts that an element will give under a certain set of experimental conditions. If the fluorescent yield value is high, the spectrometer sensitivity for that element will probably be high as well, and vice versa. The fluorescent yield takes a value from around unity for the high-atomic-number elements to as little as 0.01 for the low-atomic-number elements such as Na, Mg, and Al. It is mainly for this reason that the sensitivity of the X-ray spectrometric technique is rather poor for the elements of very low atomic number.

Most commercially available X-ray spectrometers have a range from about 0.2 to 20 Å (60–0.6 keV), which will allow measurement of the K series from fluorine ($Z = 9$) to lutetium ($Z = 71$), and for the L series from manganese ($Z = 25$) to uranium ($Z = 92$). Other line series can occur from the M and N levels, but these have little use in analytical X-ray spectrometry.

Although in principle almost any high-energy particle can be used to excite characteristic radiation from a specimen, an X-ray source offers a reasonable compromise between efficiency, stability, and cost, and almost all commercially available X-ray spectrometers use such an excitation source. Since primary (source) X-ray photons are used to excite secondary (specimen) radiation, the technique is referred to as X-ray fluorescence spectrometry.

IV. X-RAY FLUORESCENCE SPECTROMETRY

There are many types of X-ray fluorescence spectrometer available on the market today, but most of these fall roughly into two categories: wavelength-dispersive and energy-dispersive instruments. The wavelength-dispersive system was introduced commercially in the early 1950s and since then has developed into a widely accepted analytical tool. Probably around 11,000 or so such instruments have been supplied commercially, roughly half of these in the United States. Energy-dispersive

spectrometers became commercially available in the early 1970s, and today there are around 2500 units in use. The analytical chemist has available a wide range of instruments for the qualitative and quantitative analysis of multielement samples and in the choice of technique will generally consider such factors as sensitivity, speed, accuracy, cost, and range of applicability.

All conventional X-ray spectrometers comprise three basic parts: the primary source unit, the spectrometer itself, and the measuring electronics. The primary source unit consists of a very stable high-voltage generator, capable of providing up to around 3 kW of power at a potential of typically 60–80 kV, plus a sealed X-ray tube. The sealed X-ray tube has an anode of chromium, rhodium, tungsten, silver, gold, or molybdenum and delivers an intense source of continuous radiation, which then impinges on the analyzed specimen, where characteristic radiation is generated. The actual spectrometer itself consists of a specimen holder support, a primary collimator, an analyzing crystal, and a detector. The geometric arrangement of these components is shown in Fig. 4a, and a line drawing of an actual spectrometer is shown in Fig. 4b. A portion of

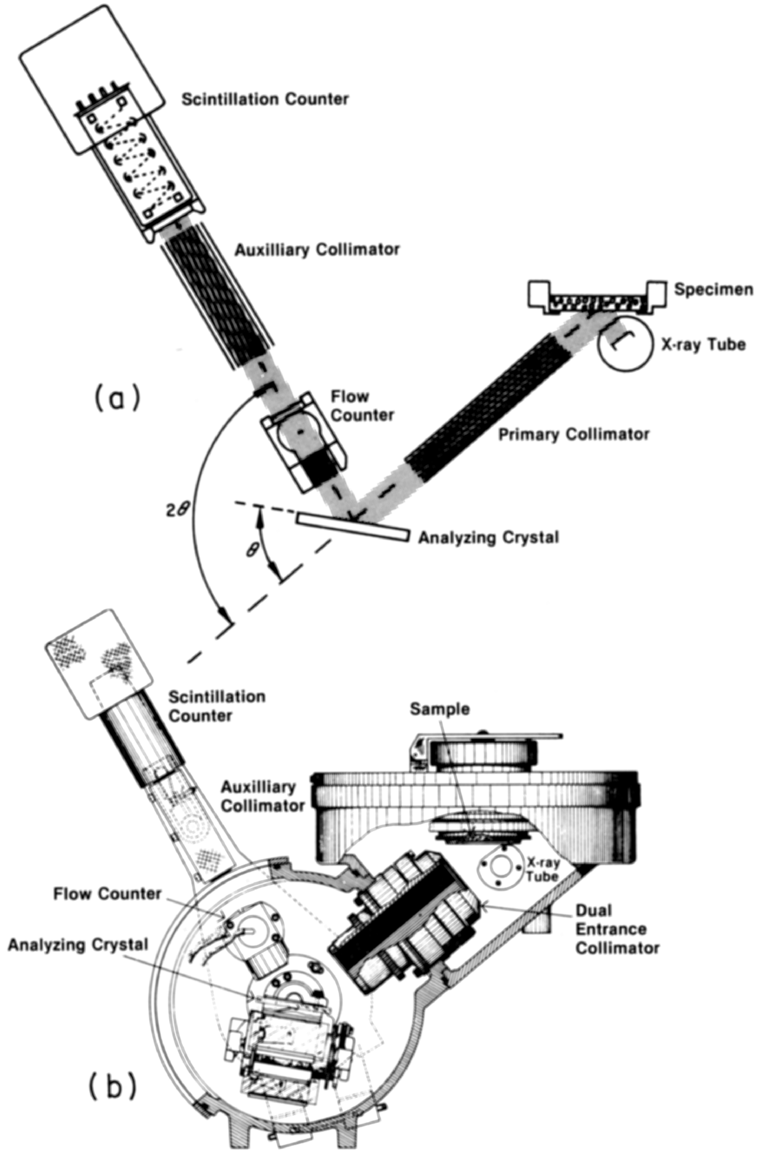


FIGURE 4 The wavelength-dispersive spectrometer. (a) The component parts of the wavelength-dispersive spectrometer. (b) The same components as they would appear in an actual commercial spectrometer.

the characteristic "fluorescence" radiation from the specimen is passed via a collimator or slit onto the surface of an analyzing crystal, where individual wavelengths are diffracted to the detector in accordance with the Bragg law. A goniometer is used to maintain the required θ–2θ relationship between crystal and detector. Typically, six or so different analyzing crystals are provided in this type of spectrometer, giving the operator a wide range of choice of dispersion conditions. In general, the smaller the d spacing of the crystal, the better the separation of the lines but the smaller the wavelength range that can be covered.

A tandem detector system is typically employed, composed of a gas flow counter and a scintillation counter. This is used to convert the diffracted characteristic photons into voltage pulses, which are integrated and displayed as a measure of the characteristic line intensity. The gas flow counter is ideal for the measurement of the longer wavelengths, and the scintillation counter is best for the short wavelengths.

The output from a wavelength-dispersive spectrometer may be either analog or digital. For qualitative work an analog output is traditionally used, and in this instance a rate meter is used to integrate the pulses over short time intervals, typically of the order of a second or so. The output from the rate meter is fed to an x/t recorder, which scans at a speed that is synchronously coupled with the goniometer scan speed. The recorder thus displays an intensity–time diagram, which becomes an intensity–2θ diagram. Tables are then used to interpret the wavelengths. For quantitative work it is more convenient to employ digital counting, and a timer–scaler combination is provided that will allow pulses to be integrated over a period of several tens of seconds and then displayed as count or count rate.

The intensities of the characteristic lines can be used to estimate the concentrations of the appropriate elements. Characteristic line intensities are related to a number of instrumental parameters, such as source intensity, diffracting efficiency of the analyzing crystal, and detector efficiency. These factors can be calibrated under a given set of experimental parameters to provide a measure of the sensitivity of the spectrometer to the element in question. The precision of an intensity measured with a modern spectrometer is of the order of one-tenth of a percent. Before this intensity can be used to calculate the element concentration, certain systematic errors from interelement effects must be corrected for. There are many standard algorithms that can be used for this process, and most automated spectrometers will allow the automatic application of these correction procedures along with the final calculation of the elemental concentrations. Accuracies of the order of a few tenths of a percent can be obtained where these procedures are correctly applied.

Most modern wavelength-dispersive spectrometers are also controlled in some way by a minicomputer or microprocessor and by use of specimen changers are capable of very high specimen throughput. Once they are set up, the spectrometers will run virtually unattended for several hours.

Like the wavelength-dispersive spectrometer, the energy-dispersive spectrometer also consist of the three basic units—excitation source, spectrometer, and detection system. In this case, however, the detector itself acts as the dispersion agent. The detector is typically a lithium-drifted silicon, "Si(Li)," detector, which is a proportional detector of high intrinsic resolution. The Si(Li) detector diode serves as a solid-state version of the gas flow detector in the wavelength-dispersive system, but with a detector gain of unity. When an X-ray photon is stopped by the detector, a cloud of ionization is generated in the form of electron–hole pairs. The number of electron–hole pairs created—that is, the total electric charge released—is proportional to the energy of the incident X-ray photon. The charge is swept from the diode by a high voltage applied across it. A preamplifier is responsible for collecting this charge on a feedback capacitor to produce a voltage pulse proportional to the energy of original X-ray photon. Thus when a range of photon energies is incident upon the detector, an equivalent range of voltage pulses is produced as a detector output. A multichannel analyzer is used to sort the arriving pulses to produce a histogram representation of the X-ray energy spectrum.

The output from an energy-dispersive spectrometer is generally displayed on a CRT, and Fig. 5 shows such an output obtained from a geological sample using a molybdenum-anode X-ray tube. The X-ray tube is sometimes fitted with a thin transmission filter, the function of which is to lower the background. In the figure, two spectra were acquired, the first with the molybdenum tube alone and second with the tube fitted with a molybdenum filter. The counting time in both instances was 400 s and the effect of the filter on the background is clearly seen. Fifteen elements have been identified. The operator is able to display the contents of the various channels dynamically as an energy spectrum. Provision is generally made to allow zooming in on portions of the spectrum of special interest, to overlay spectra, to subtract background, and so on, in a rather interactive manner. As in the case of the modern wavelength-dispersive systems, nearly all energy-dispersive spectrometers will incorporate some form of minicomputer that is available for spectral stripping, peak identification, quantitative analysis, and a host of other useful functions.

As has been previously mentioned, within the two major categories of X-ray spectrometers specified, there is a wide diversity of instruments available. The major differences

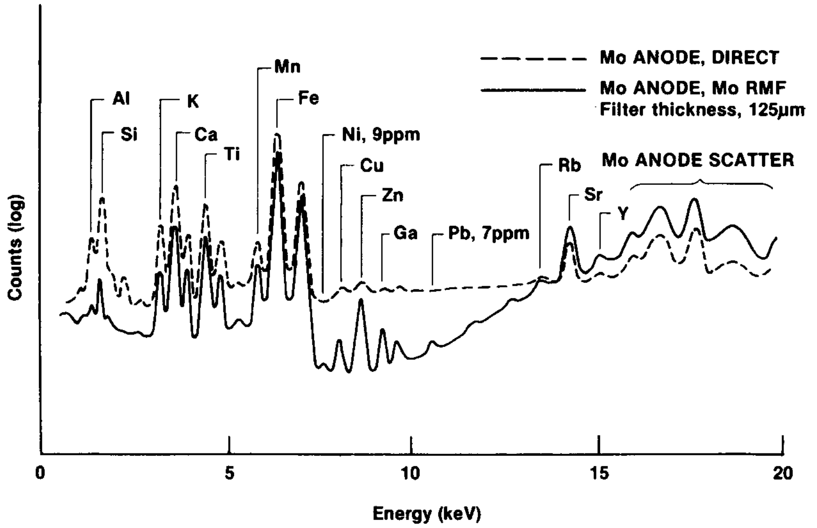


FIGURE 5 Geological sample, 400-s analysis. A multichannel analyzer output from a rock sample, using a 400-s counting time. Two spectra are compared: dashed line is direct excitation using an Mo-anode tube; dotted line is the same but using an Mo filter between tube and sample.

generally lie in the type of source used for excitation, the number of elements that are measurable at one time, the speed at which they collect data, and finally the price range. All of the instruments are, in principle at least, capable of measuring all elements in the periodic classification from $Z = 9$ (F) and upward. Most can be fitted with multisample handling facilities, and all can be automated by use of minicomputers. All are capable of precisions of the order of a few tenths of a percent, and all have sensitivities down to the low parts per million (ppm) level. As far as the analyst is concerned, they differ only in their speed, cost, and number of elements measurable at the same time.

V. APPLICATION OF X-RAY SPECTROMETRIC METHODS

The great flexibility and range of the various types of X-ray fluorescence spectrometer, coupled with their high sensitivity and good inherent precision, make them ideal for quantitative analysis. Single-channel wavelength-dispersive spectrometers are typically employed for both routine and nonroutine analysis of a wide range of products, including ferrous and nonferrous alloys, oils, slags and sinters, ores and minerals, thin films, and so on. These systems are very flexible but relative to the multichannel spectrometers are somewhat slow. The multichannel wavelength-dispersive instruments are used solely for routine, high-throughput analyses where the great need is for fast, accurate analysis, but where flexibility is of no importance. Energy-dispersive spectrometers have the great advantage of being able to display information on all elements at the same time. They lack somewhat in resolution compared to the wavelength-dispersive systems but find great application in quality control, trouble-shooting problems, and so on. They have been particularly effective in scrap alloy sorting, in forensic science, and in the provision of elemental data to supplement X-ray powder diffraction data.

Like all instrumental methods of analysis, the high precision can be translated into high accuracy only if the various systematic errors in the analysis process are taken care of. The precision of a well-designed X-ray spectrometer is typically of the order of ~0.01%; the major source of this random error is the X-ray source, that is, the high-voltage generator plus the X-ray tube. In addition, there is an error arising from the statistics of the actual counting process. The production of X rays is a random process. Thus, in any experiment a number of counts N are taken by counting a certain count rate R for a period of time t, and there will be a random error σ associated with a measured value of N, this being equal to $\sqrt{N}$:

$$\sigma = \sqrt{N} = \sqrt{R \times t} \tag{4}$$

Since the measured value of N is time-dependent, the counting error can be reduced to the level of other random errors (e.g., the source error) simply by taking a suitable value of t. Equation (4) can be extended for the case where background is significant or where multiple measurements are employed.

Systematic errors in quantitative X-ray spectrometry arise mainly from absorption- and specimen-related phenomena, much as in X-ray powder diffraction, except that in the spectrometry case they are much more complicated. This is because a single characteristic wavelength is used

in the diffraction experiment, whereas many wavelengths are involved in spectroscopy. This means, for example, that the equivalent effect expression given in Eq. (1) carries an integral sign. Although these so-called matrix effects are somewhat complicated, many excellent methods have been developed for handling them. The advent of the minicomputer-controlled spectrometer has done much to enhance the application of these correction procedures, and today in most cases one is able to quantify most elements in the periodic table of atomic number 9 (F) and upward to an accuracy of a few tenths of a percent. The areas of application of the X-ray fluorescence technique now cover almost all areas of inorganic analysis.

VI. TRACE ANALYSIS BY X-RAY FLUORESCENCE SPECTROMETRY

The X-ray fluorescence (XRF) method is particularly applicable to the qualitative and quantitative analysis of low concentrations of elements in a wide range of samples, as well as allowing the analysis of elements at higher concentrations in limited quantities of materials. The sensitivity m of the XRF method is expressed in terms of the intensity of the measured wavelength per unit concentration, expressed in counts per second per percent. Figure 6 shows the approximate counting rate that would be obtained from 1% of the indicated element and shows that

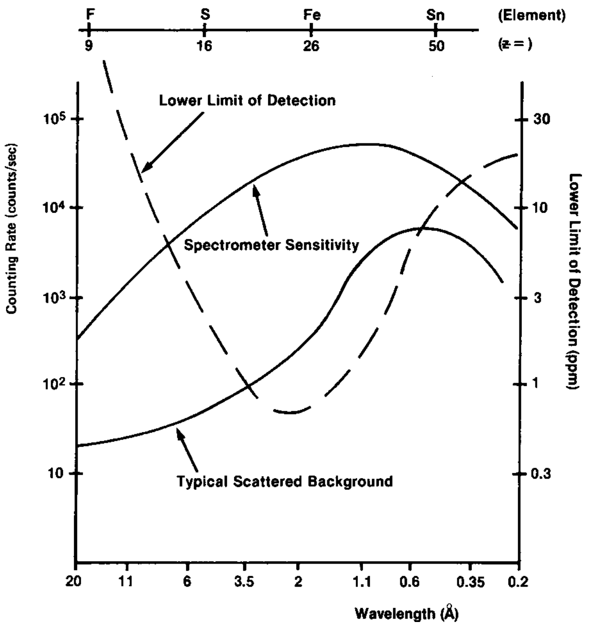


FIGURE 6 Sensitivity, background, and lower limit of detection curves. Plots of spectrometer sensitivity in counts per second per percent, background counting rates in counts per second, and lower limits of detection in parts per million, as a function of wavelength.

the sensitivity varies by about three orders of magnitude over the normal range of 0.2 to 20 Å. The lower limit of detection for a given element is related to the sensitivity for the element in question, relative to the background intensity at the measured wavelength. The lower limit of detection (LLD) is expressed as the concentration of the element equivalent to two standard deviations of the background counting rate R_b. It is generally assumed that in the low concentration region the counting error is the major source of random error, and the equation for LLD is related to the background counting time T_b in the following manner:

$$\text{LLD} = (3/m)\sqrt{R_b/T_b} \tag{5}$$

The background counting time T_b is equal to half of the total counting time T. The factor 3 in the equation arises from the fact that two measurements must be made, that is, peak and underground, which increases the two standard deviations by $\sqrt{2}$. As an example, the magnesium K_a line measured in an aluminum alloy gives a sensitivity m of about 1000 counts/s percent, over a background R_b of about 30 counts/s. In a total counting time of 100 s ($T_b = 50$ s) the LLD would equal $(3/1000) \times (30/50)^{1/2}$ or 0.0023%.

It will also be noted from the figure that not only does the sensitivity of the spectrometer vary significantly over the wavelength range of the spectrometer, but so too does the background counting rate. The background curve shown was obtained using a chromium-anode X-ray tube operated at 50 kV, using a sample of very low atomic number, that is, of high scattering power. It will be seen that the background varies by about two orders of magnitude over the range of the spectrometer. By inspection of Eq. (5), it will be seen that the detection limit will be best when the sensitivity is high and the background is low. A combination of these two factors leads to the detection-limit curve shown in the figure. The curve is approximate, since both the spectrometer sensitivity and the measured background vary with the average atomic number of the sample. Nevertheless, detection limits over most of the atomic number range lie in the low part per million range. The sensitivity of the X-ray spectrometer falls off quite dramatically toward the long-wavelength limit of the spectrometer, due mainly to low fluorescence yields and the increased influence of absorption. As a result, poorer detection limits are found at the long-wavelength extreme of the spectrometer, which corresponds to the lower atomic numbers. Thus the detection limits for elements such as fluorine and sodium are at the levels of hundredths of a percent rather than parts per million.

Even though the long-wavelength limit of the spectrometer is generally taken as about 20 Å, it is possible to extend this limit to about 50 Å by use of special techniques.

This allows the measurement of two more elements, carbon ($Z = 6$) and oxygen ($Z = 7$). The detection limits for these elements are, however, very poor and are typically of the order of 3–5%.

The irradiation area in a typical X-ray spectrometer is of the order of 5 cm^2 and the penetration depth of an average wavelength is about 20 μm. This means that even though 20 g of sample may be placed in the spectrometer, the analyzed volume is still only of the order of 50 mg. The smallest sample that will give a measurable signal above background is at leas three orders of magnitude less than this mass: ~0.05 mg, provided that the sample is spread over the full irradiation area of the spectrometer sample cup. Where this is impracticable, the smallest value that can be analyzed is increased by a factor corresponding to the ratio of the area of the beam (~5 cm^2) to the actual area of the sample. In practical terms, this sensitivity is quite sufficient for the sample masses typically encountered in the analysis of trace metals in air and water samples, and XRF methods find increasing application in this area. A second important area of materials analysis involving small samples involves the investigation of thin films, and, although the technique is limited to rather large areas (typically a few square millimeters), it does provide useful information about bulk composition of surface films.

VII. THE CRYSTALLINE STATE

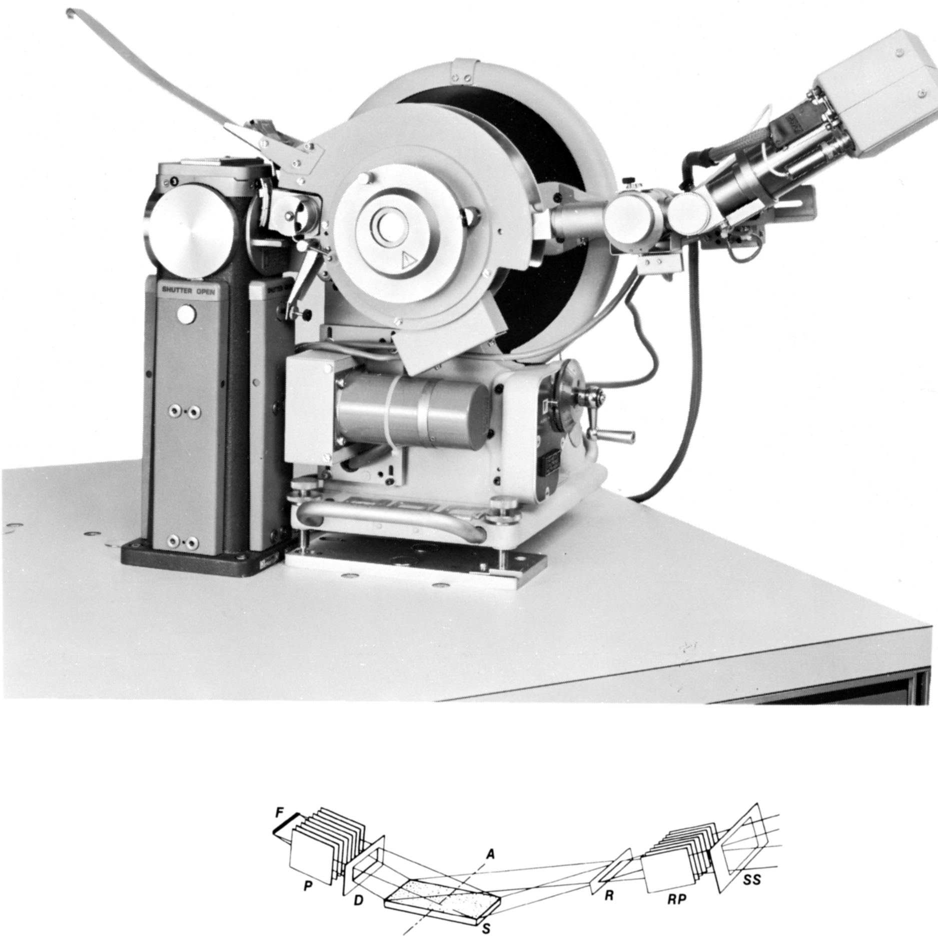
All substances are built up of individual atoms, and nearly all substances have some degree of order of periodicity in the arrangement of these atoms. A crystal can be defined as a homogeneous, anisotropic body having the natural shape of a polyhedron. In practical terms, whether a substance is homogeneous or not can only be defined by the means that are available for measuring the crystallinity. In general, the shorter the wavelength, the smaller the crystalline region that can be recognized. Even noncrystalline materials have a degree of order, and each will give some sort of a diffraction pattern. For example, glassy materials and liquids will give diffraction patterns of sorts, generally in the form of one or more broad diffuse peaks or halos. A crystalline substance has a definite form, which is retained no matter what the physical size of the crystal. A certain type of crystal can thus be defined in terms of specific physical characteristics that determine its shape. Since every ordered material is made up of a unique arrangement and number of atoms, every ordered material will give a diffraction pattern that is, to all intents and purposes, also unique.

In X-ray powder diffractometry one is generally dealing exclusively with polycrystalline materials, and the specimen does not actually have to be a powder. The diffraction method is equally applicable to bulk metals, thin films, polymer sheets, and a variety of other forms. In addition to the use of the method for qualitative and quantitative phase identification, it can also be used to determine the degree of crystallinity, that is, the dimensions of the crystalline regions in otherwise amorphous substances. Where the crystal lattice is distorted by strain, this will also manifest itself in modifications to the diffraction pattern. Where the strain is isotropic, the lines in the diffraction pattern will be displaced; where the strain is anisotropic, the lines will be broadened. Study of selected line positions and line shapes can reveal much information about stress/strain conditions within a given polycrystalline material such as a quenched steel. Study of diffraction intensity distributions occuring over the whole surface of a specimen can reveal information about the "texture" of a material. This technique is referred to as a "pole figure" study. Similar techniques can be used to determine topographic information from the surface of single crystals. Diffraction patterns can also be recorded over a wide range of temperatures from liquid helium up to 2000°C. These low- and high-temperature experiments are invaluable for the study of phase transformations. However, by far the greatest single use of the powder technique is for phase identification of inorganic, mineral, and organic solids.

VIII. X-RAY POWDER DIFFRACTOMETRY

The instrumentation that is used for powder diffraction measurements has not changed much from the instruments developed in the late 1940s. The major difference found in modern instrumentation is the use of the minicomputer for control, data acquisition, and data processing. Figure 7a shows a photograph of a typical vertical powder diffractometer system, and Fig. 7b illustrates the geometry of the system. This geometric arrangement is known as the Bragg–Brentano parafocusing system and is typified by a diverging beam from a line source F, falling onto the specimen S, being diffracted and passing through a receiving slit R to the detector. Distances FA and AR are equal. The amount of divergence is determined by the effective focal width of the source and the aperture of the divergence slit D. Axial divergence is controlled by two sets of parallel-plate collimators (Soller slits) P and RP placed between focus and specimen and between specimen and scatter slit, respectively.

Use of the narrower divergence slit will give a smaller specimen coverage at a given diffraction angle, thus allowing the attainment of lower diffraction angles where the specimen has a larger apparent surface (thus larger values of d are attainable). This is achieved, however, only



(b)

FIGURE 7 (a) The vertical diffractometer. A Philips PW1050 vertical goniometer system. (b) Geometry of the Bragg–Brentano diffractometer. The geometrical layout of a typical diffractometer system showing source F, Soller slits P and RP, sample S, divergence slit D, and receiving slit R. The axis of the goniometer is at A.

at the expense of intensity loss. Choice of the divergence slit, plus its matched scatter slit, is thus governed by the angular range to be covered. The decision as to whether or not the slit size should be increased at a given angle will be determined by the available intensity. A photon detector, typically a scintillation detector, is placed behind the scatter slit and converts the diffracted X-ray photons into voltage pulses. These pulses may be integrated in a rate meter to give an analog signal on an x/t recorder. By synchronizing the scanning speed of the goniometer with the recorder, a plot of degrees 2θ versus intensity, called the diffractogram, is obtained. A timer/scaler is also provided for quantitative work and is used to obtain a measure of the integrated peak intensity of a selected line(s) from each analyte phase in the specimen. A diffracted beam monochromator may also be used in order to improve signal-to-noise characteristics. The output from the diffractometer is a "powder diagram," essentially a plot of intensity as a function of diffraction angle, which may be in the form of a strip chart or a hard copy from a computer graphics terminal.

The powder method derives its name from the fact that the specimen is typically in the form of a microcrystalline powder, although, as has been indicated, any material that is made up of an ordered array of atoms will give a diffraction pattern. The possibility of using a diffraction pattern as a means of phase identification was recognized by 1935, but it was not until the late 1930s that a systematic means of unscrambling the superimposed diffraction patterns was proposed by Hanawalt, Frevel, and Rinn. Their technique was based on the use of a file of single-phase patterns, characterized in the first stage by their three strongest reflections, and a search technique based on matching strong lines in the unknown pattern with these standard pattern lines. A potential match was then confirmed by a check using the full pattern in question. The identified pattern

was then subtracted from the experimental pattern, and the procedure was repeated on the residue pattern until all lines were identified.

Manual techniques for this "search/matching" process have changed little over the years. In the hands of experts manual search/matching is an extremely powerful tool, but for the less experienced user it can be rather time-consuming. Typically 2–4 h may be required for the complete identification of a four-phase mixture. A growing complication is that the file of standard patterns increases by about 2000 each year and currently stands at about 46,000 entries. Most manual methods of search/matching that are used to identify phases in an unknown mixture are similar to that shown in Fig. 8. The three strongest lines in the pattern are used to locate potential matches in the index of standards. Each time a potential candidate is found, a match is made with the complete pattern. If all lines agree, a phase confirmation is assumed and the lines for the match are subtracted from the original pattern. This process is repeated until all significant lines in the pattern are identified.

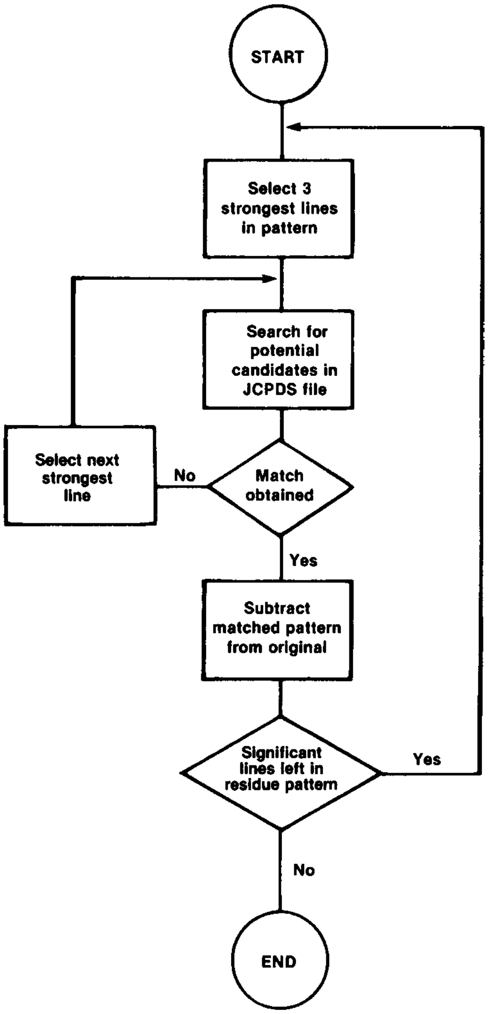


FIGURE 8 Analytical approach to multiphase diffraction. A flow chart showing the major steps in performing a typical qualitative analysis by X-ray diffraction.

The two basic parameters being used in this search/match process are the d values calculated from the measured 2θ values in the diffractogram and the relative intensities of the lines in the pattern. Whereas the d value can be accurately measured (with an accuracy of better than 0.5% in routine analysis), the intensities are rather unreliable by comparison and can be subject to error, sometimes running into tens of percent. Because of the unreliable nature of the measured intensities, any search procedure based on the selection of lines by their intensities must be used with caution.

The responsibility for the maintenance of the Powder Data File lies with the International Centre for Diffraction Data (JCPDS), which is a nonprofit organization located in Swarthmore, Pennsylvania. This group is made up of a staff of permanent officers along with a number of academic and industrial scientists who are active in the field of X-ray powder diffractometry. The Powder Data File is a unique assembly of good-quality single-phase patterns and is used by thousands of chemists, geologists, and materials scientists. The automation of the search/match process promises to make the use of the file even more widespread, since this should do much to relieve much of the tedium associated with manual search/matching.

IX. APPLICATION OF X-RAY DIFFRACTION METHODS

Of all of the methods available to the analytical chemist, only X-ray diffraction is capable of providing general-purpose qualitative and quantitative information on the presence of phases (e.g., compounds) in an unknown mixture. Although it is true that techniques such as differential thermal analysis will provide some information on specific phase systems under certain circumstances, such methods could not be classified as general-purpose. As described in the previous section, a diffraction pattern is characteristic of the atomic arrangement within a given phase, and to this extent it can act as a "finger-print" of that particular phase. Thus by use of the JCPDS powder data file a series of potential matches can be obtained. A complication in the application of this method for the analysis of multiphase materials is that the patterns are superimposed on top of one another and consequently there may be uncertainties as to which lines belong to which phases. As a result of this, for many years effective use of the powder diffraction method was restricted to a rather

limited number of experts in the field. In more recent years the minicomputer has played an enormous role in increasing the effectiveness of searching, so that now relatively inexperienced users are also able to analyze even rather complex mixtures.

The whole of the JCPDS Powder Data File can now be stored onto a 10-megabyte disk, and suitable search match programs can be run on an average-size minicomputer. As an example, Fig. 9 shows a flow chart of one such commercially available program, typical of what is currently available from most instrument manufacturers. The diffractogram is reduced to a list of d spacings and intensity values by means of a peak searching routine, and this d/I list is then submitted for file search/matching. In order to simplify the search procedure further, the JCPDS has subdivided the file of standard patterns into mineral, inorganic, organic, metals, and alloys. Where these subfile restrictions are known, the total search procedure can be restricted to the appropriate subfile(s), giving a significant reduction in the time required for the search/match process. A file of common phases, sometimes called frequently encountered phases, is also available, and a presearch is made of this file, plus a prematch of selected preconceived possibilities. As an example, if one were analyzing a specimen obtained by aspirating air through a filter paper, and the X-ray spectrometer indicated significant concentrations of silicon and iron, likely candidates to search for in the diffraction pattern would be SiO_2, Fe_2O_3, and $FeSiO_3$.

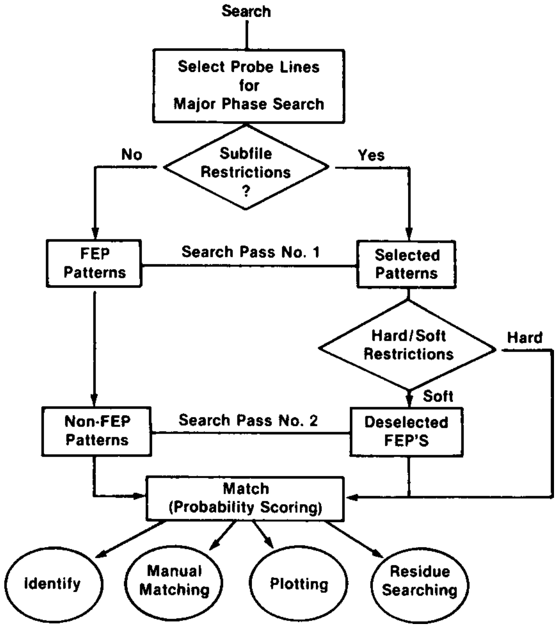


FIGURE 9 Search/match flow chart. A flow chart showing the prescreening processes used in computer search/matching of diffraction data. FEP refers to "frequently encountered physics."

In practice it is rather unusual to analyze specimens in which nothing at all is known; generally one has access to information about, for example, sample history and environment. In addition, other analytical information can be employed where available. As an example, in the routine analytical laboratory it is common to find the X-ray diffractometer and the X-ray spectrometer being used to give complementary information. This type of information can be used to impose what are called "hard" and "soft" restrictions to the search. The restictions are used to generate a subset of the total data file, and that subset is searched pattern by pattern, matching the d/I list of the unknown sample to those known phases in the subset. In the example given in Fig. 9, a probability matching scheme is used for this latter process. The final output from the program is a list of possible matches, each given with a score representing the certainty of the match. In addition, some method of graphically representing the original pattern and possible matched candidates is also provided.

Although the X-ray powder diffraction technique is invaluable for qualitative phase identification, it is neither a sensitive nor a rapid means of analysis. The minimum detectable limit found by routine qualitative procedures would be of the order of several percent, compared, for example, with a few parts per million in X-ray fluorescence. A complete analysis would take several hours to complete, although this will of course depend to a very large extent on the experience of the analyst and the complexity of the problem in hand. The method is applicable to almost any crystalline material, whether inorganic or organic.

Once the presence of a phase has been established in a given specimen, one can, at least in principle, determine how much of that phase is present by use of the intensities of one or more diffraction lines from the phase. However, as has been pointed out previously, it may be difficult to obtain an accurate value for these intensities. The intensities of the diffraction peaks are subject to a variety of random and systematic errors, and the errors fall roughly into three categories:

1. Structure-dependent: a function of atomic size and atomic arrangement, plus some dependence on the scattering angle and temperature.
2. Instrument-dependent: a function of diffractometer conditions, source power, slit widths, detector efficiency, etc.
3. Specimen-dependent: a function of phase composition, specimen absorption, and particle size, distribution, and orientation.

For a given phase or selection of phases, all structure-dependent terms are fixed and in this instance have no

influence on the quantitative procedure. If one calibrates the diffractometer with a sample of the pure phase of interest and then uses the same conditions for the analysis of the unknown mixture, the systematic errors that are instrument-dependent are constant and only the random errors associated with counting statistics need be considered. However, the specimen-dependent factors remain. The most important of these are the effects of size, distribution, and orientation of particles and the effects of absorption.

The absorption effect has already been referred to in Section II. Clearly, in a multiphase mixture, different phases will absorb the diffracted photons by different amounts. As an example, the mass absorption coefficient for CuK_α radiation is 308 cm/g for iron, but only 61 cm/g for silicon. Thus iron atoms are five times more efficient than silicon atoms in absorbing CuK_α photons. There is a variety of standard procedures for correcting for the absorption problem, of which by far the most common is the use of the internal standard. In this method a standard phase is chosen that has about the same mass absorption coefficient as the analyte phase, and a weighed amount of this material is added to the unknown sample. The relative intensities of lines from the analyte phase and the internal standard phase are then used to estimate the relative concentrations of internal standard and analyte phases. The relative sensitivity of the diffractometer for these two phases is determined by a separate experiment. Other procedures are available for the analysis of complex mixtures, but these are beyond the scope of this particular work. For further information the reader is referred to specific texts dealing with the X-ray powder method.

The handling of particle problems is unfortunately much more complex. As has been previously stated, the powder method requires a specimen that is randomly oriented since the geometry of the system requires that an equal number of crystallites be in the correct position (i.e., orientation) to diffract at any diffraction angle where the goniometer happens to be. Where particles lie in a preferred orientation there will be more particles available to diffract at the angle corresponding to this orientation, and, equally important, fewer particles available to diffract at other diffraction angles. The overall effect is to enhance some intensities and to diminish others. Thus, the intensities are dependent on particle distribution and orientation. Some materials, just by virtue of their crystal habit, may become preferentially oriented during sample preparation. As an example, mica, being a rather "platy" material, will prefer to stack one plate on top of another rather than take up a random orientation. The overall effect of preferred orientation can introduce errors that vary from insignificant tens of percent. Careful specimen preparation is always called for, and this may include grinding, sieving, spray-drying, and a host of other techniques.

The areas of application of quantiative X-ray powder diffraction are many and varied. Many hundreds of analysts are using this technique on a daily basis. Some of the more common applications would include ore and mineral analysis, quality control of rutile/anatase mixtures, retained austenite in steels, determination of phases in airborne particulates, various thin-film applications, study of catalysts, and analysis of cements. The current state of the art in the quantitative analysis of multiphase materials is that accuracies of the order of a percent or so can be obtained in those cases where the particle orientation effect either is nonexistent or has been adequately compensated for.

X. COMBINED DIFFRACTION AND SPECTROSCOPIC TECHNIQUES

Like all analytical techniques, both X-ray powder diffraction and X-ray fluorescence have their advantages and disadvantages. Table I indicates the more important features of the two techniques. In terms of the range of application, X-ray fluorescence analysis allows the quantitation of all elements in the periodic table from fluorine (atomic number 9) upward. Accuracies of a few tenths of a percent are possible, and elements are detectable in most cases to the low parts per million level. X-ray powder diffractometry is applicable to any ordered (crystalline) material, and although it is much less accurate or sensitive than the fluorescence method, it is almost unique in its ability to differentiate phases.

The techniques differ widely in terms of their speed of analysis. The modern multichannel wavelength-dispersive spectrometer is able to produce data from 20–30 elements in less than 1 min. An energy-dispersive spectrometer, or even a relatively simple wavelength-dispersive spectrometer, is able to perform a full qualitative analysis on an unknown sample in less than 30 min. The diffraction technique, on the other hand, can take as much as an

TABLE I Features of X-Ray Spectrometry and Diffraction

Feature	Spectrometry	Diffractometry
Range	$Z > 5$	All ordered materials
Speed	Seconds to minutes	Minutes to hours
Precision	0.1%	0.25%
Accuracy	0.1–1.0%	0.5–5%
Sensitivity	Low ppm	0.1–2%
Cost	$50,000–250,000	$25,000–125,000

order of magnitude longer than this. Even with the most sophisticated computer-controlled powder diffractometers available today, diffraction experiments are invariably very time-consuming. A similar difference also exists in the sensitivities of the two methods. Whereas the fluorescence method is able to measure signals from as little as 1 ppm of a given element, the diffraction technique is often hard put to measure 1%.

The spectrometry method also outperforms the diffraction technique both in precision and accuracy, perhaps by a factor of two to six. The diffraction technique is at best a rather insensitive, slow technique, giving somewhat poor quantitative accuracy. On the other hand, it will be realized that the information given by the X-ray diffraction method is unique, and no other technique is able to provide such data. This is not true of the X-ray fluorescence method, since there are many other techniques available to the analytical chemist today for the quantitation of elements. Atomic absorption, inductively coupled plasma emission, and solid-source mass spectrometry are all examples of competing methods.

The fluorescence and diffraction techniques are to a large extent complementary, since one allows accurate quantitation of elements to be made and the other allows qualitative and semi-quantitative estimations to be made of the way in which the matrix elements are combined to make up the phases in the specimen.

Both the simultaneous wavelength-dispersive spectrometer and the energy-dispersive spectrometers lend themselves admirably to the qualitative analysis of materials. As was shown in Eq. (3), there is a simple relationship between the wavelength or energy of a characteristic X-ray photon and the atomic number of the element from which the characteristic emission line occurs. It was also stated in Section II that each element will emit a number of characteristic lines within a given series (K, L, M, etc.). Thus by measuring the wavelengths, or energies, of a given series of lines from an unknown material, the atomic numbers of the excited elements can be established. Because the characteristic X-ray spectra are so simple, the actual process of allocating atomic numbers to the emission lines is a relatively simple process and the chance of making a gross error is rather small. The procedures for the qualitative analysis of multiphase materials with the X-ray powder diffractometer are a much more complex business. There are after all only 100 or so elements, and within the range of the conventional spectrometer each element gives, on an average, only half a dozen lines. In diffraction, on the other hand, there are as many as several million possible compounds, each of which can give on an average 50 or so lines. However, a combination of spectrometric and diffraction data can often be used to greatly simplify what would otherwise be a complex and time-consuming analysis.

SEE ALSO THE FOLLOWING ARTICLES

CRYSTALLOGRAPHY • LASERS, X-RAY • X-RAY PHOTOELECTRON SPECTROSCOPY

BIBLIOGRAPHY

Automated powder diffractometry, new dimensions in instrumentation and analytical software. (1983). *Norelco Reporter* (special issue) **30**.

Barrett, C. S., and Gilfrich, J. V. (1990). "Advances in X-Ray Analysis," Vol. 33. Plenum, New York.

Dzubay, T. G., ed. (1977). "X-Ray Fluorescence Analysis of Environmental Samples," Ann Arbor Science, Ann Arbor, Mich.

Holy, V., Pietsch, U., and Baumbach, T. (1998). "High-Resolution X-Ray Scattering from the Films and Multilayers," Springer Tracts in Modern Physics, Vol. 149, Springer-Verlag, Berlin.

Jenkins, R. (1984). X-Ray technology. *In* "Kirk-Othmer Encyclopedia of Chemical Technology," Vol. 24, 3rd ed., pp. 678–708. Wiley, London.

Jenkins, R., Gould, R. W., and Gedcke, D. A. (1981). "Quantitative X-Ray Spectrometry," Marcel Dekker, New York.

Kikkert, J. N. (1984). X-ray spectrometry in the eighties. *Norelco Reporter* **31,** 20–25.

Klug, H. P., and Alexander, L. E. (1974). "X-Ray Diffraction Procedures," 2nd ed. Wiley, New York.

Michelte, A., and Pfauntsch, S., eds. (1996). "One Hundred Years of X-Rays," John Wiley & Sons, New York.

Nagel, D. J. (1982). Potential characteristics and applications of X-ray lasers. *In* "Advances in X-Ray Spectroscopy" (M. Bonnelle and C. Mande, eds.), pp. 371–410. Pergamon, Oxford.

Siegbahn, K., ed. (1967). "Atomic Molecular and Solid State Structure Studied by Means of Electron Spectroscopy," Almquist, Uppsala.

Turcu, I. C. E., and Dance, J. B. (1998). "X-Rays from Laser Plasmas: Generation and Applications," John Wiley & Sons, New York.

Van Griener, R. E., section ed. (2000). "X-ray spectrometry." *In* "Encyclopedia of Analytical Chemistry" (R. A. Meyers, ed.), John Wiley & Sons, Chichester, U.K., pp. 13269–13443.

Zevin, L. S., Kimmel, G., and Mureinik, I. (1995). "Quantitative X-Ray Diffractometry," Springer-Verlag, Berlin.

X-Ray Photoelectron Spectroscopy

Charles C. Chusuei
D. Wayne Goodman
Texas A&M University

I. Fundamentals of XPS
II. Instrumentation
III. Analysis of Electronic and Chemical Structure of Atoms, Molecules
IV. Examples

GLOSSARY

Binding energy The energy (measured in eV) of the exiting *photoelectron* produced by the photoionization process of core level (first shell) electrons, which contain discrete chemical information. Peaks arising from energy level of various orbitals allow for qualitative oxidation state identification.

Core level shift Precise photoelectron binding energy peak positions of the peak centers denote the chemical oxidation and/or electronic states of orbitals from which the photoelectrons emanate. Minute variations in the binding energy arising from differing electronic environments of the ejected photoelectrons are manifested as shifts in the peak position (also referred to as *chemical shifts*).

Fermi level A reference point (taken as zero eV) with which binding energies of photoelectrons are measured.

Inelastic mean free path The average distance that a particle at a given energy (photoelectrons) can travel along a trajectory between inelastic collisions within a solid.

Intra- and extra-atomic relaxation The contraction of outer electronic orbitals toward the nucleus resulting from core hole vacancies following the photoemission process. The outgoing photoelectrons can be screened by either valence electrons within the atom (intra-atomic) or from local environment electrons just outside the atom (extra-atomic). These effects contribute to a surface analyte's core-level binding energy position observed in the X-ray photoelectron spectra.

Monolayer A one-molecule thick quantity of adsorbate on the surface.

Shake-up satellites Excitation of valence electrons (that accompanies relaxation processes) to an unfilled level at higher binding energy. The loss of kinetic energy of the outgoing photoelectron into a discrete

state appears as a peak along with the main core-level photoionization.

Surface sensitivity The ability to probe the topmost (on the order of angstroms) atomic layers of a solid.

Work function The energy required to remove an electron from the solid. This energy level above the Fermi reference is known as the *vacuum level.*

X-ray-excited Auger emission A secondary electron emission process that follows the photoionization and appears as a peak in the X-ray photoelectron spectrum. After the initial photoemission, an upper level valence electron relaxes into the vacant core-level state, followed by an ejection of another electron in the valence level.

X-RAY photoelectron spectroscopy (XPS) is an important and widely used surface analysis method in a many fields of study in physics and chemistry (e.g., microelectronics, heterogeneous catalysis, environmental geochemistry, etc.). The technique probes the energy distribution of electrons ejected from solids via irradiation by X-rays and the photoelectric effect; the electrons contain information regarding chemical oxidation state, electronic structure, and atomic composition of the analyte being studied. Thus, surface composition as well as the electronic environment can nondestructively (in many cases) be determined. X-ray photoelectron spectroscopy is also useful for quantitative analysis, capable of probing ultrathin layers of material (0.1% of a monolayer). Since the photoelectrons analyzed emanate only from the topmost atomic layers of the solid surface being studied ($\leq$100 Å), the technique is an invaluable tool for studying interfacial phenomena at the solid-solid and solid-gas boundaries.

I. FUNDAMENTALS OF XPS

A. The Photoelectric Effect and Secondary Processes

X-ray photoelectron spectroscopy (XPS) operates on the principle of the photoelectric effect, which occurs via a primary excitation process brought about by X-ray-irradiation producing electrons (*photoelectrons*) of discrete energy, containing chemical information regarding the surface analyte. It should be noted that X-rays are only one of many types of excitation sources that can be used to induce emission of electrons for analysis. X-ray photoelectron (XP) spectral peaks (generated by the photoelectrons) are named according to the orbital ($l = 0, 1, 2, 3 \ldots$ denoted as s, p, d, f. . .) and spin ($s = \pm 1/2$) quantum numbers of the core levels from which they emanate. The total momentum of the photoelectrons ($J = l \pm s$) is included in the nomenclature of a measured XP spectral peak (e.g., Ca $2p_{3/2}$ where $l + s = 1 + 1/2 = 3/2$).

The technique is sensitive to all elements except H and He and has a detection limit of about 0.1% of a monolayer. There is little systematic overlap of the spectral lines between elements albeit some does occur, such as Ca 2p/Au 3d, Pt 4f/Al 2p, O 1s, Sb 3d and Al 2s, 2p/Cu 3s, 3p. Difficulties arising from spectral overlap are easily overcome via examination of additional transitions (excitations) that can arise from the same element. In well-calibrated XPS systems, the precision of the quantitation measurements is typically within ±5%; thus, XPS is well-suited for both quantitative and qualitative elemental analysis.

In the photoemission process, an atom absorbs a photon of a known energy (hν) resulting in an ejection of a core level electron, which is detected and its energy measured. Figure 1 shows an energy level diagram relevant for binding energy measurements of a photoelectron escaping from the solid (sample specimen) and into the electron spectrometer. The conducting specimen is in electrical contact with the spectrometer housing so that both the sample and spectrometer have a common reference for measuring electron energy, called the *Fermi level*, E_F. Incoming photons (with energy $h\nu$) create a photoelectron with kinetic energy, E_k^1 relative to the *vacuum level*, E_v, of the sample. The kinetic electron energy at the sample surface, E_k^1, is determined from the kinetic energy of the electron, E_k, measured inside the spectrometer, from the relation:

$$E_k = E_k^1 - (\phi_{spec} - \phi_s) \quad (1)$$

where ϕ_{spec} and ϕ_s are *work functions* (energies required to remove electrons from the Fermi to vacuum levels) of the spectrometer and sample, respectively. From Fig. 1, it is evident that the binding energy of analyte from an electrically conducting sample may be obtained from the relation:

$$h\nu = E_b + E_k + \phi_{spec} \quad (2)$$

where $h\nu$ = photon energy from and an X-ray source and E_b = binding energy. Notably, ϕ_s is not involved in the measurement of the E_k that is "seen" by the spectrometer. In practice, the E_b of the incoming electrons is computed from the measured E_k from Eq. 2. Photoelectrons are generated via ejection from the solid surface upon excitation of core level orbitals by a photon source:

$$A + h\nu \rightarrow A^{+*} + e^- \text{ (photoelectron)} \quad (3)$$

where A is a neutral atom or molecule, A^{+*} is the excited ion and e^- is the ejected photoelectron. The photoionization along with competitive secondary processes are illustrated in Fig. 2. XPS owes its relatively nondestructive nature to the fact that only the ejection of electrons is

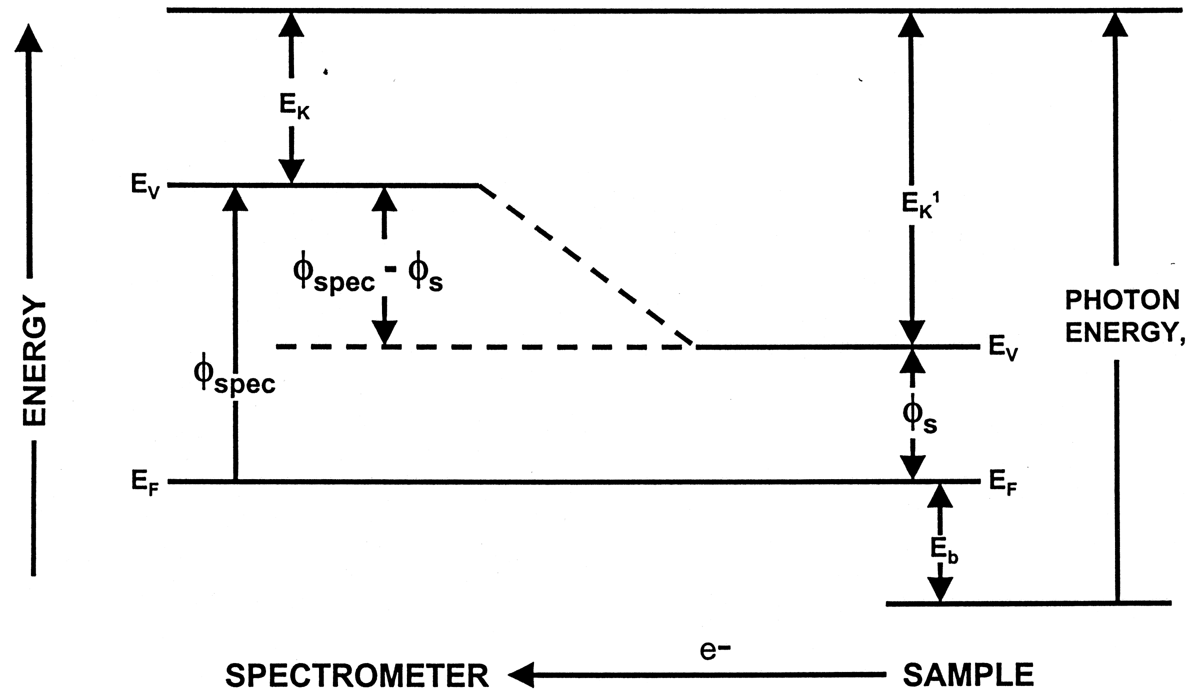


FIGURE 1 Energy level schematic for XPS binding energy measurements.

required for analysis. Unlike other techniques of elemental analysis, the atomic nuclei being examined remain unchanged during electron spectroscopic measurements. It should be noted that although some samples are sensitive to decomposition from exposure to the X-ray source, this is not generally intrinsic to the analysis technique. The photoionization is accompanied by two secondary emissions, characterized by either a photon emission resulting in *X-ray fluorescence*:

$$A^{+*} \rightarrow A^{+} + h\nu' \text{ (X-ray photon)} \qquad (4)$$

or an ejection of a valence level electron, called an *Auger* electron, whose kinetic energy is independent of the photon energy:

$$A^{+*} \rightarrow A^{++} + e^{-} \text{ (Auger electron)} \qquad (5)$$

The initial photoionization is a two-step process, which can produce either Auger electrons or X-ray photons (i.e., fluorescence). X-ray fluorescence and Auger emissions are competitive processes. The Auger effect dominates for low-energy photoionization processes (about 1 kV) while X-ray emission dominates at high energies (about 10 keV). These secondary emissions are used for other surface analytical methods: X-ray fluorescence (XFS) and Auger electron (AES) spectroscopies. Since conventional XP spectrometers utilize a relatively low photon energy (i.e., Mg Kα at $h\nu = 1253.6$ eV; Al K$\alpha = 1486.6$ eV), contributions from fluorescence in the XP spectra are generally negligible. On the other hand, Auger transitions that are excited by the impinging X-ray photons commonly appear in XPS. X-ray-excited Auger electron spectroscopy (XAES) is innately a part of XPS, but is typically considered a separate technique because of the ways in which the data are analyzed. Figure 3 shows an XPS *survey scan* (0–1000 eV) of Ni deposited onto an H-ZSM-5 zeolite powder catalyst (DeGussa; consisting of mostly SiO_2) obtained using a Mg Kα source operated at 300 W. The XP

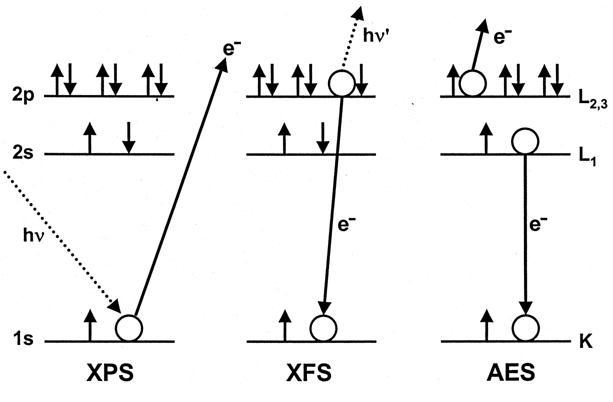


FIGURE 2 Diagrams depicting the photoemission process for XPS along with secondary emission processes (XFS and AES). [Reproduced with permission from the *Journal of Chemical Education*, Vol. 61, No. 6, 1984, p. 483; Copyright © 1984, Division of Chemical Education.]

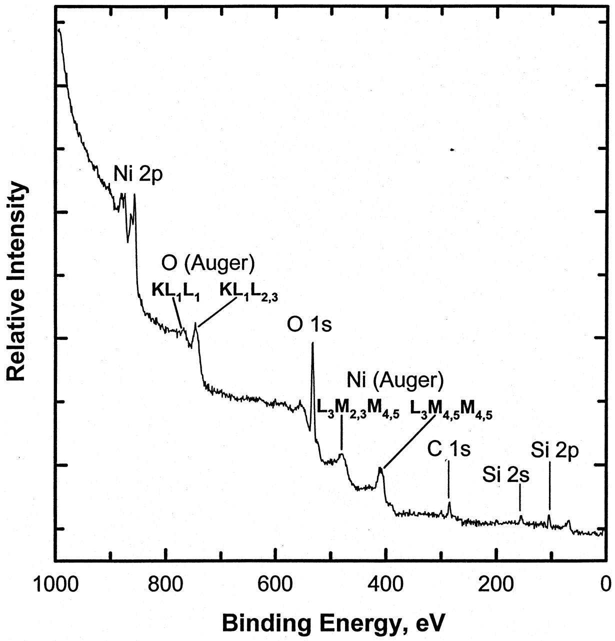


FIGURE 3 XP survey scan of Ni/H-ZSM-5 catalyst (DeGussa) containing O KLL and Ni LMM Auger transitions.

spectrum is a plot of the counting rate versus the binding energy of the detected photoelectrons. Other XPS core levels that are readily observed in the survey scan include those of the Ni 2p, C 1s, O 1s, and Si 2p orbitals. Different XPS transitions arising from various orbitals from a single element can appear at multiple peak positions in the survey scan. For example, Si has photoelectron transitions from the 2s and 2p levels that appear at about 150 and 100 eV (Fig. 3), respectively. Note that the high binding energy side of the XPS peak (e.g. O 1s, C 1s) has a higher background; this increase in count is due to partial loss of kinetic energy of the ejected electrons from inelastic collisions as it travels through the solid. Counts emanating from these partial losses have a reduced kinetic energy and would thus show up at higher binding energy in the XP spectrum (Eq. 2). The high background observed (Fig. 3) at about 900 eV extending up to the bombardment photon excitation energy is due to *bremmstralung* (German for "braking radiation") that is generated from the X-ray source.

The Auger effect is characterized by an upper level valence electron relaxation into the vacant core-level state (after the initial photoionization), followed by an ejection of another electron in the valence level. In the nomenclature of the Auger effect, for example, for the $KL_1L_{2,3}$ transition in Fig. 2, the first shell corresponds to the core level in which the initial vacancy was created (K) either via photoemission from XPS or electron impact bombardment from an electron beam, the second letter (shell) refers to the upper level valence electron that relaxes to fill the vacancy (L_1), and finally the third shell is the Auger electron that is ejected into the vacuum level ($L_{2,3}$) and then detected and measured via the spectrometer. Four groups of XAES lines are generated by conventional Mg Kα and Al Kα X-rays: (1) the KLL [Na, Mg, O, F, Mn, Fe, Co], (2) the LMM [Cu, Ni, Zn, Ga, Ge, As, Se], (3) the MNN [Ag, Cd, In, Sn, Sb, Te, Ru, Rh, Pd, I, Xe, Cs, Ba], and (4) the NOO [Au, Hg, Tl, Pb, Bi]. The X-ray-excited Auger peaks in Fig. 3, denoted by the arrows at the O KL_1L_1, O $KL_1L_{2,3}$, Ni $LM_{2,3}M_{4,5}$, and Ni $L_3M_{4,5}M_{4,5}$ transitions, typically have broader full-width half-maxima (fwhm) than the XPS lines (as compared to the fwhm of the O 1s, Ni $2p_{1/2}$, and $2p_{3/2}$ photoelectron emissions). Although too broad to be useful as indicators of chemical shifts, the Auger electron kinetic energy is independent of the primary excitation energy and can be used as reference points for detecting small chemical oxidation state changes (see Section III.A).

B. Properties of Solids and Surface Sensitivity

In order for electron spectroscopy to be effective, the ejected photoelectron has to "escape" from the solid, into the vacuum level, and finally into the electron energy analyzer and detector. Only electrons located near the top surface can escape without loss in kinetic energy because of the high probability of inelastic scattering inside the solid. Thus, XPS has severe limitations for studying bulk material. Surface compositions of materials can substantially differ from that inside the bulk due to segregation or surface contamination. On the other hand, the *surface sensitivity* of the technique (i.e., ability to probe the topmost atomic layers) is advantageous for studying interfaces and analyzing surface coverages that are only fractions of a monolayer thick. Thus XPS is valuable for the investigation of molecular-level adsorbates on solid surfaces. Figure 4 shows a "universal curve" plot of the *inelastic mean free path* (IMFP) versus the kinetic energy of the electron. The IMFP is defined as average distance that a particle (photoelectrons) can travel along a trajectory between inelastic collisions within a solid. The larger the kinetic energy, the longer the IMFP. In this particular plot, the IMFP of various elements (Ag, Au, Mo, Be, P, C, W) is given in angstrom units. A broad minimum IMFP is evident at 4–10 Å corresponding to a 10–500 eV electron kinetic energy. While photoionization occurs up to a few microns below the surface, only the electrons at the first few tens of angstroms exit the solid without energy loss and provide most of the intensity for XPS. As a rule of thumb, approximately 95% of the XPS signal arises from depths less than or equal to three times the IMFP.

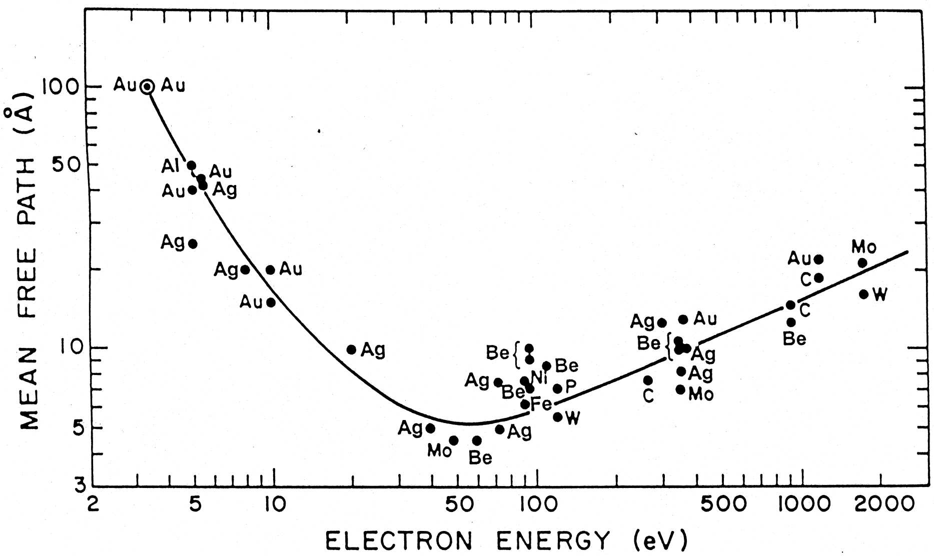


FIGURE 4 Universal curve plot of electron IMFPs of various elements. [Reprinted with permission from Somorjai, G. A. (1981). "Chemistry in Two Dimensions: Surfaces," p. 41, Cornell University Press, Ithaca, NY.]

II. INSTRUMENTATION

A. Ultrahigh Vacuum

A vacuum environment is necessary for XPS experiments to be undertaken for three reasons. First, low pressure is required to insure a sufficient inelastic mean free path for the photoelectrons to escape the solid and reach the electron detector without being dissipated via inelastic collisions. From the kinetic theory of gases, the mean free path of the gas molecule (analogous to the IMFP concept of photoelectrons traveling through a solid), λ', can be determined from

$$\lambda' = \frac{1}{d_0^2 n \sqrt{2}} \tag{6}$$

where d_0 = molecular diameter of the gas and n = density of molecules. In order for the technique to be effective, λ' must at least span the distance between the sample and analyzer, typically <1 m. Hence, in order to avoid dissipation via inelastic collisions of the escaping photoelectrons (or Auger electrons in the case of XAES), pressures of 10^{-5} Torr or lower are required. Second, even lower pressures (about 10^{-7} Torr; in the *high vacuum* pressure regime) are required to operate the X-ray anode and filaments without damage. Finally, in surface science experiments, low pressures are needed to minimize surface contamination from interfering with analysis. From the kinetic theory of gases, higher pressures accompany increased surface bombardment with gas molecules. Assuming a unity sticking coefficient, a 10^{-6} Torr pressure for 1 sec (defined as the unit, *Langmuir*) will produce a 1.0 monolayer of contamination. Thus, at 10^{-7} Torr the surface is contaminated after 10 sec, at 10^{-8} Torr the surface is contaminated after 100 sec, etc. Typically, pressures in the *ultrahigh vacuum* (UHV) regime (defined as 1×10^{-9} Torr or lower) are employed to maximize the analysis time before the surface is contaminated. To remove contamination, most systems are equipped with a means of *in situ* sample cleaning by *ion sputtering* and/or sample heating. Sputtering involves the bombardment of the substrate with a beam of an inert gas (such as He, Ar) to remove impurities via momentum transfer. Especially in the case of single crystal refractory metal substrates, sample heating for cleaning is typically accomplished by passing current through the specimen or electron beam (e-beam) bombardment. Temperatures as high as 2000 K can be readily achieved via e-beam heating.

B. Radiation Sources

In selecting an elemental source suitable for producing useful XP spectra, several factors need to be considered: (1) the energy resolution of the X-rays; (2) energy of the photons that are produced; and (3) ease of application of the material to an anode surface. Y Mζ ($h\nu$ = 132.3 eV; fwhm = 0.47 eV) and Zr Mζ ($h\nu$ = 152.4 eV; fwhm = 0.77 eV) lines produce photon energies that are too low in energy to be effective since they excite a narrow range of photoelectrons, which limits the number of elements that can be analyzed. Cu Kα ($h\nu$ = 8048 eV; fwhm = 2.5 eV) has a sufficiently large excitation line to cover the full range of elements; however, its natural line width is too large (ideal fwhm <1.0 eV); this would limit

chemical oxidation state analysis due to poor spectral resolution. Si Kα produces a sufficiently high photon energy and narrow resolution to be a suitable X-ray source ($h\nu = 1739.5$ eV; fwhm = 1.00 eV). However, its insulating properties and difficulties associated with applying the material onto an anode surface make its use impractical. The Al Kα ($h\nu = 1486.6$ eV; fwhm = 0.85 eV) and Mg Kα ($h\nu = 1253.6$ eV; fwhm = 0.70 eV) lines suffer none of these limitations and are almost universally used in laboratory XPS instruments. Both have sufficient photon energy to excite characteristic lines of a wide range of elements and their natural line widths are sufficiently narrow to allow differentiation between multiple oxidation states. Both anodes produce XP spectra where overlap of photoelectron and X-ray-excited Auger lines can occur in certain photoelectron binding energy regions, complicating analysis. For this reason, most commercial systems produce X-ray sources with a dual anode system (e.g., Mg Kα/Al Kα). The XAES/XPS signal overlap of one anode is oftentimes not present when switching to the other anode for use. Both anodes can be used, respectively, during an experiment to eliminate these "blind spots."

It should be noted that while the Mg or Al radiation source is typically labeled with "Kα" for the sake of brevity, this unmonochromatized, "natural" line (spectrum) radiation is more complex. The X-ray lines (of each source) are an unresolved doublet, K$\alpha_{1,2}$, that have satellite features, K$\alpha_{3,4}$ and Kβ. Excitations from the K$\alpha_{3,4}$ and Kβ appear as satellites along side the photoelectron peak(s) produced by the K$\alpha_{1,2}$ and lead to spectral misinterpretation. K$\alpha_{3,4}$ excitations are $\sim$10 eV on the low binding energy side of the main core-level peak and is approximately one-tenth of the K$\alpha_{1,2}$ intensity; Kβ excitations appear $\sim$70 eV on the lower binding energy side. Removal of the satellites can be achieved via monochromatization of the X-ray source. These satellite intensities to some degree in nonmonochromatized systems are eliminated by using an aluminum foil X-ray window (approximately 8 μm thick) to reduce the bremsstrahlung radiation from the X-ray source.

Mg Kα and Al Kα radiation are typically excited with electron energy that is about 10 times greater than their respective photon energies (e.g., 15 kV, 300 W) for efficient production of photons. To insure optimum sensitivity and detection, the analyte surface is brought as close to the X-ray source as possible. Since a relatively large amount of energy is used to generate the X-ray source within a compact area (in which the X-ray anode is held), the anode must be kept sufficiently cool to avoid damage (to both the sample and anode). Deionized cooling water is typically used to transfer heat away from the anode (in thermal contact with a copper block). Conducting ions from the coolant fluid are removed chiefly for two reasons: (1) to avoid current generation that may be exhibited as artifacts in the XP spectrum and (2) as an electrical safety precaution due to the fact that the anode is typically operated at high voltage.

The drawback of using discrete line sources (as mentioned above) is that photoionization cross sections for different core levels excited by a fixed photon energy differ markedly; some cross sections will be at a maximum while others are at a minimum for the same line energy. Moreover, as discussed above there are severe limitations in the number of choices for other fixed line sources. The use of a continuously tunable source to circumvent this obstacle is available in the form of *synchrotron* radiation. In synchrotron, electrons are accelerated to near-relativistic speeds via pulsed magnetic fields around a circular ring, known as a *torus*, giving rise to X-ray irradiation. These electrons emit light in a continuous spectrum, having a shape dependent upon the radius of curvature and electron energy. Since the radiation is concentrated into a narrow cone tangential to the electron orbit, it can be easily passed into a monochromator for energy selection. Photon energies can be monochromatized to very high resolution in this manner. Another distinct advantage of using synchrotron radiation is that surface sensitivity can be tuned via adjusting the photon energy close to that of the photoionization threshold. Since the IMFP is proportional to the kinetic energy of the photoelectron, kinetic energies can be tuned to about 25–100 eV. The photoelectron signals thus obtained would have high surface sensitivity.

Synchrotron XPS systems are immense in size. The torus is typically tens of meters in diameter, providing energy and current outputs of several GeV and 1A, respectively. Stronger S/N is readily seen (as compared with single photon line XPS instruments) throughout the entire equivalent conventional XPS binding energy regime. In addition to the tunability of the X-ray source, the synchrotron radiation is 100% plane-polarized, pulsed, and has the advantage of allowing for time-resolved experiments to be performed. Due to its considerable expense in construction, repair, maintenance, and operation as well as the large space needed to house the apparatus, synchrotron XPS systems tend to be nationally owned and shared among numerous outside researchers. Thus, there are severely restricted time allowances available for experiments. Synchrotron instruments, due to their complexity, also tend to be out of operation for maintenance and repairs much more frequently than laboratory XPS instruments. Nevertheless, to many scientists the gains offered by a synchrotron source (not available to conventional laboratory XP spectrometers) warrant the inconvenience associated with its use.

C. Binding Energy Calibration

Since the XPS binding energy scaling may be compressed or expanded, points both at high and low binding energy bounds are needed in order to calibrate the entire binding energy span. The calibration procedure needs to be undertaken periodically to insure accurate binding energy measurements. This is typically accomplished with sputter-cleaned metal foils that are free of contaminants, such as carbon and oxygen. Energy positions of well-documented lines include Au $4f_{7/2}$ (84.0 eV; low) and Cu $2p_{3/2}$ (932.7 eV; high) are typically used. Binding energy peak centers of their respective photoelectron transitions are obtained via high-resolution narrow scans. The distance (i.e., scaling) between the high and low binding energy points is a parameter known as the *scale factor*, which can be controlled using the instrumental software. In practice, the scale factor is first adjusted and then followed by changing the *work function* [ϕ in Eq. (2)] to shift the entire binding energy scale so that the measured peaks match those of the standards. The work function adjustment oftentimes will change the scale factor and vice versa; thus, both variables should be changed taking into account the results of their interplay. This procedure thus requires multiple sets of XPS scans of high and low binding energy core levels. Figure 5 shows an XP survey scan of Au and Cu foils juxtaposed alongside each other; both the Au $4f_{7/2}$ and Cu $2p_{3/2}$ orbitals are thus readily accessible for instrumental calibration. Accompanying XAES Cu LMM features are also seen in the spectrum.

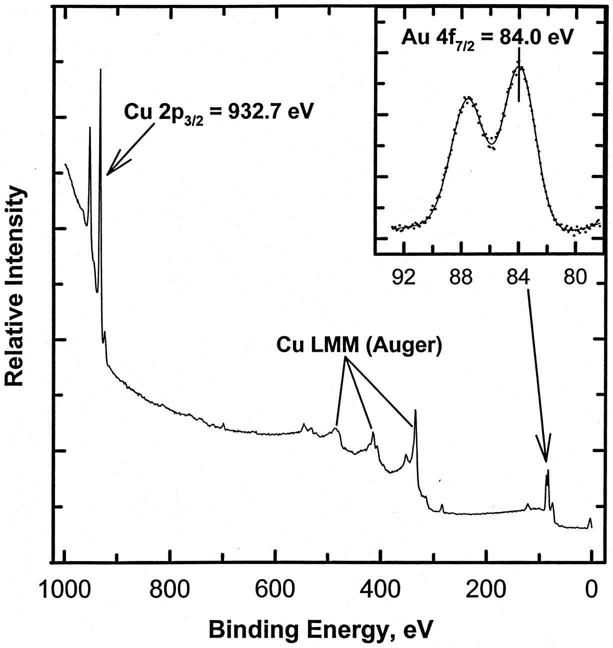


FIGURE 5 Sputter-cleaned Cu and Au foils (juxtaposed together on sample specimen holder). Binding energy peak positions of the Cu $2p_{3/2}$ and Au $4f_{7/2}$ levels serve as calibration standards. The inset shows an expanded view of the Au 4f orbital intensities.

D. Sample Charging

Many sample specimens have electrically insulating properties that would affect the XPS binding energy measurements and impede spectral interpretation. Sample *charging* occurs as a result of an accumulation of positive charge (as electrons are ejected away from the solid during photoemission) onto the surface when grounding of the sample (that would otherwise restore electrical neutrality) is impaired. For the nonconducting surface, electrons cannot return to the surface easily and are lost faster than they return. After a steady state is reached, a positive charge develops due to deficiency in electron density. Photoelectrons from the analyte would thus be ejected with a decreased kinetic energy (higher binding energy) along with broadening of the peak fwhm.

Corrections for charging can be accomplished by either (1) the use of reference standards or (2) an electron flood gun to neutralize the positive surface charge. One common practice is to evaporate Au *dots* onto the surface (after detailed high-resolution scans) and use the Au $4f_{7/2}$ level as the reference. In the case where the sample specimen has been exposed to atmospheric pressure, the C 1s line at 284.7 ± 0.2 eV denoting *adventitious carbon* (from hydrocarbons prevalent in the air) can be used. It should be noted that this type of correction may not adjust for changes in the scale factor due to *differential* (inhomogeneous) charging, i.e., some areas of substrate have more charge than others. In this instance, the charge correction for a high binding energy region of the XP spectrum may differ from that of a low one. A flood gun may be used instead to irradiate electrons to the sample specimen in order to neutralize charging. However, care should be taken in its use since many materials, especially organic substances, are sensitive to electron beam induced decomposition.

E. Energy Analyzers

X-ray photoelectron spectroscopy analysis allows for identification of various elements as well as differentiation of various chemical oxidation states. In order to accomplish these tasks, the energy resolution must be sufficiently high so that peaks from different chemical oxidation states can be deconvoluted. Furthermore, the resolution should remain relatively constant for all elemental transitions, i.e., the *absolute* resolution, ΔE, (defined by the fwhm) must remain the same. If the linewidths are to be matched by the absolute resolution at the maximum photon energies

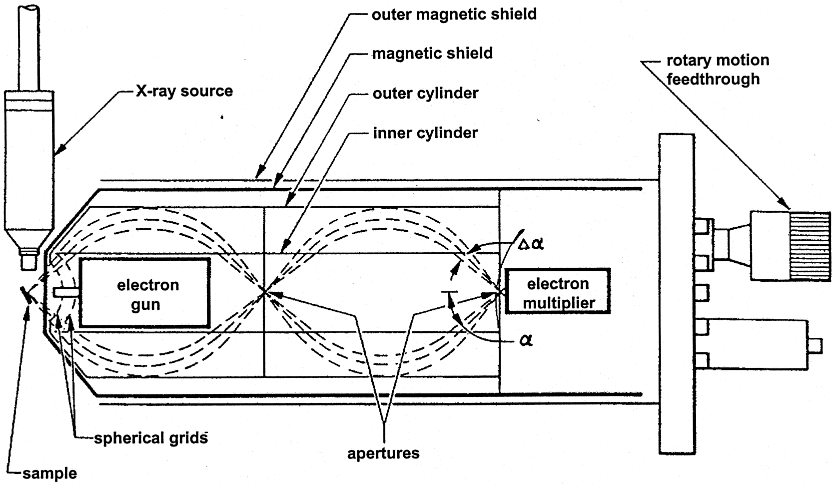


FIGURE 6 Double-pass cylindrical mirror analyzer for XPS and AES analysis. [Reproduced with permission from P. W. Palmberg, *Journal of Vacuum Science and Technology A*, Vol. 12, No. 1, 1975, p. 380; Copyright © 1975, American Institute of Physics.]

(e.g., fwhm = 0.70 eV for Mg Kα), a *relative* resolution (defined as $R = \Delta E/E_0$ where E_0 is the kinetic energy of the peak position) of (0.70 eV ÷ 1253.6 eV =) 6×10^{-4} would be required. Attaining this resolution is feasible but the data obtained herein would suffer unacceptable losses in sensitivity. In practice, the kinetic energies of the incoming photoelectrons (toward the detector) are retarded to a preselected analyzer energy and kept fixed during data acquisition. Thus, for a pass energy of 50 eV, the relative resolution needed would be (0.70 eV ÷ 50 eV =) 1×10^{-2}, which can be easily achieved with minimal loss in sensitivity. Decreasing the pass energy effectively increases resolution. In the commonly used double-pass cylindrical mirror analyzer (CMA) setup (Fig. 6), two spherical grids at the front of the first stage are used to retard electrons (from the sample) entering the analyzer (and finally into the electron multiplier) when used in the XPS mode. The X-ray source is usually positioned as close as possible to the sample in this configuration. An electron gun is housed inside the CMA to excite Auger transitions. In some applications, the electron gun can also be modified for use as a flood gun to correct for charging. As a precaution, however, electron beam currents should not exceed 1×10^{-8} A since high currents can damage the electron multiplier. In an Auger electron spectrometer, only a single CMA is used to optimize the luminosity of the ejected Auger electrons. In this design for XPS, two CMAs are connected together in series to increase resolution; the exit aperture of the first stage is the entrance aperture to the second. Thus, the double-pass CMA evolved (i.e., from AES analysis) for use in XPS systems. In front of the first stage are spherical grids used to retard incoming photoelectrons at the preselected constant pass energy (by applying voltages to the inner and outer cylinders). Inner and outer magnetic shields depicted prevent external magnetic fields from altering the photoelectron flight path. The entrance angle into the CMA is fixed at $42.3 \pm 6°$. When the analyzer is used in the AES mode (best performed without retardation), these grids are grounded via a rotary motion feedthrough adjustment.

Unlike the CMA which was developed for use for both AES and XPS, the *concentric hemispherical analyzer* (CHA) was fabricated specifically for XPS since its inception. The CHA (Fig. 7) consists of two concentrically positioned hemispheres with inner (R_1) and outer (R_2) radii; ϕ and r are the angular and radial coordinates of the photoelectrons (with kinetic energy E_0) entering the analyzer at an entrance angle α. These spheres have negative and positive potentials, respectively, when voltages (V_0 and V) are applied across them. The median equipotentials between the hemispheres is found at radius, $R_0 = (R_1 + R_2)/2$. Similar to the CMA, it is customary to preretard the photoelectrons as they enter the analyzer. The angle of acceptance into the analyzer is adjustable. In the case where the angle (ϕ) between the entrance and exit aperture is 180°, as the incoming photoelectron travel across R_0 with kinetic energy, E_0, the deflecting potential, eV, varies with E_0 by

$$\text{eV} = E_0\left(\frac{R_2}{R_1} - \frac{R_1}{R_2}\right) \quad (7)$$

In order to optimize energy resolution, α can be adjusted so that $\alpha \approx \frac{w}{2R_0}$ where w = the slit width. Similar to the CMA system, photoelectrons are retarded to preselected pass energies via planar grids across the entrance slit as they enter

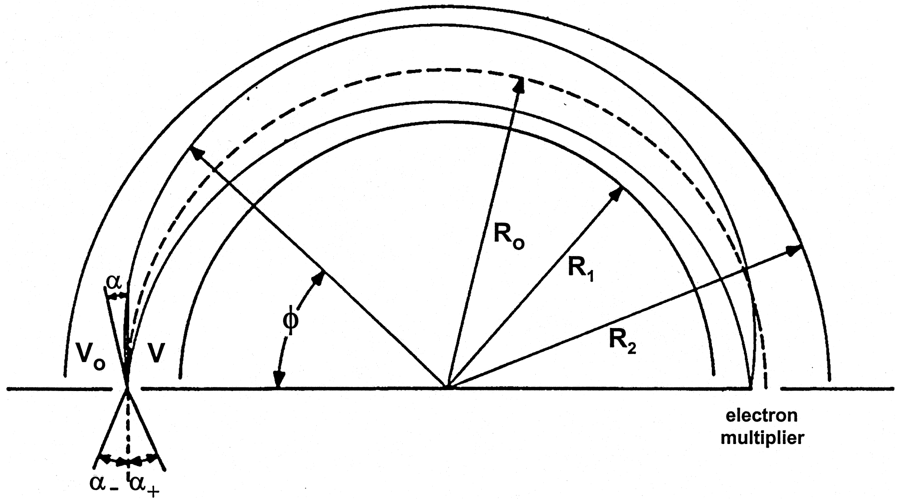


FIGURE 7 Concentric hemispherical analyzer (CHA) diagram. [Reproduced with permission from D. Roy and J.-D. Carette, *Canadian Journal of Physics*, Vol. 49, p. 2141, 1971; Copyright © 1971, NRC Research Press.]

the CHA. The CHA has the added option of the use of the lens system for improvement of overall efficiency of detection. The entrance slit to the CHA is narrow and one-dimensional, which gives the added advantage of *angle-resolved* analysis capabilities used for depth profiling the specimen.

F. Depth Profiling, Angle-Resolved Analysis

The sample specimen is rotated normal to the direction of the slit entrance to the CHA in order to adjust the surface-sensitivity of the analyte being sampled. Figure 8 shows a schematic depicting the relation between the IMFP, λ, and the average escape depth of the photoelectrons:

$$d = \lambda \sin \alpha \tag{8}$$

where d = the depth of the photoelectrons escaping from the solid and α = the *take-off* angle (angle between the flight of the photoelectrons from the solid and the surface itself). In this particular scenario, a more surface-sensitive depth is sampled, $d_1(\alpha_1 < 90°)$, in "A" as compared to $d_2\,(\alpha_2 \approx 90°)$ in "B"; decreasing the take-off angle increases surface sensitivity. In the case where the surface being analyzed consists of multiple layers (perhaps topmost layers have differing chemical oxidation states as compared to the bulk), a series of XPS scans performed varying the take-off angle would enable the surface and bulk components to be differentiated. By sampling the multiple layers via changing the surface specificity of the analysis in this manner, a *depth profile* of the top surface layers can be obtained. It should be noted that this type of analysis is only effective in the case in which the surface layers or thin films have thicknesses at the same order of magnitude as the IMFP, provided that the surface is flat. Angle-resolved analysis is not available with the CMA system due to its large solid angle, which impairs its ability to detect photoelectrons emanating from a single direction.

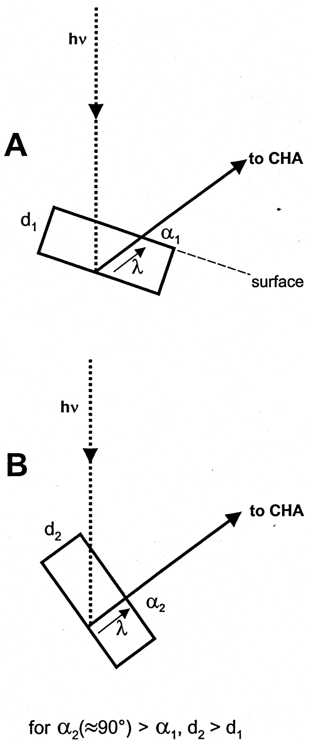


FIGURE 8 Angle-resolved analysis for depth profiling. Schematics show how surface sensitivity can be adjusted via changing the angle (α) of the path of ejected photoelectrons with respect to the CHA.

The single-dimensionality of the entrance slit also allows the added benefit for the CHA to be fitted with an input lens system.

III. ANALYSIS OF ELECTRONIC AND CHEMICAL STRUCTURE OF ATOMS, MOLECULES

A. Chemical Oxidation State Determination

Chemical shift data are typically obtained from high-resolution scans of smaller binding energy regions to determine precise peak center positions (also referred to as *narrow scans*). Small differences in the binding energy peak position (typically 0–3 eV) are attributed to chemical shifts arising from different electronic environments of the ejected photoelectrons. Variations in the number of valence electrons and the type of bonds they form (between neighboring atoms) influence the binding energy of the ejected photoelectrons, thus giving rise to chemical shifts exhibited in differences in the peak centers. Table I shows the chemical shifts (in eV) measured with respect to the zero oxidation state. Note that for each element depicted, as the oxidation number increases (becoming more positive), binding energy increases due to greater attraction of the nucleus to a core electron by the relative absence of outer valence electrons. When one of these outer shell electrons is removed, the effective charge sensed by the core electron increases (arising from a deficiency of electron density). Conversely, in the case where the attraction of the nucleus for a core electron is diminished by the presence of outer electrons, photoelectrons ejected from an electron-rich environment will exhibit a lower binding energy. Multiple oxidation states can appear in a given narrow scan. In order to make meaningful peak assignments to these different states, issues regarding instrumental resolution need to be addressed. In theory, the ΔE of a single oxidation state observed in an XP spectrum is a convolution of three components:

$$\Delta E = \sqrt{\Delta E_n^2 + \Delta E_p^2 + \Delta E_a^2} \tag{9}$$

where ΔE_n is the natural width of the core level, ΔE_p is the width of the photon source, and ΔE_a is the width from the analyzer resolution. Both ΔE_p and ΔE_a can be controlled experimentally by choosing an optimal photon source and varying instrumental settings of the analyzer, respectively. The photon line widths, ΔE_p, for standard Mg Kα and Al Kα sources are 0.70 and 0.85 eV, respectively, and are fixed values. ΔE_n, however, depends on the uncertainty of the core hole lifetime of the ionized state in the photoemission process. From the uncertainty principle, this natural line width can be expressed as:

$$\Delta E_n = \frac{\mathrm{h}}{\tau} = \frac{4.1 \times 10^{-15}}{\tau}\ \mathrm{eV} \tag{10}$$

where h = Planck's constant (eV · sec) and τ (sec) is the lifetime of the photoemission process. The fastest and slowest lifetimes vary between 10^{-15} to 10^{-13} sec, resulting in line widths with a lower and upper limits of 0.04 and 4 eV, respectively. Thus, based on the natural line widths, these fwhm values represent the upper and lower bounds of a single chemical oxidation state. In practice, the *absolute resolution* (ΔE) is obtained from by measuring the fwhm of an "ideal" photoelectron peak (from the analyte of interest) that is free from broadening effects or experimental artifacts (having a symmetrical Gaussian shape) and has the narrowest peak width in the XPS survey scan. The ΔE obtained can serve as a guide for deconvolution of multiple peaks when curvefitting spectra. For example, Fig. 9 shows an O 1s XPS peak separated into contributions from hydroxyl oxygen and the TiO_2 metal oxide oxygen. The curvefitting was not allowed to deviate from a 2.0–2.5 eV fwhm. The XP spectrum "A" shows the O 1s signal after a TiO_2 substrate was exposed to aqueous solution for

TABLE I Chemical Shifts as a Function of Oxidation State[a]

Element	Oxidation state									
	−2	−1	0	+1	+2	+3	+4	+5	+6	+7
Nitrogen 1s	—	*0[b]	—	+4.5[c]	—	+5.1	—	+8.0	—	—
Sulfur 1s	−2.0	—	*0	—	—	—	+4.5	—	+5.8	—
Chlorine 2p	—	*0	—	—	—	+3.8	—	+7.1	—	+9.5
Copper 1s	—	—	*0	+0.7	+4.4	—	—	—	—	—
Iodine 4s	—	*0	—	—	—	—	+5.3	—	—	+6.5
Europium 3d	—	—	—	—	*0	+9.6	—	—	—	—

[a] All shifts are measured in eV relative to the oxidation states denoted by the asterisks (*).
[b] Arbitrary zero measurement, end nitrogen in NaN_3.
[c] Middle nitrogen in NaN_3.
Reprinted with permission from Hercules, D. M. (1970). *Anal. Chem.* **42,** 28A; Copyright © 1970, American Chemical Society.

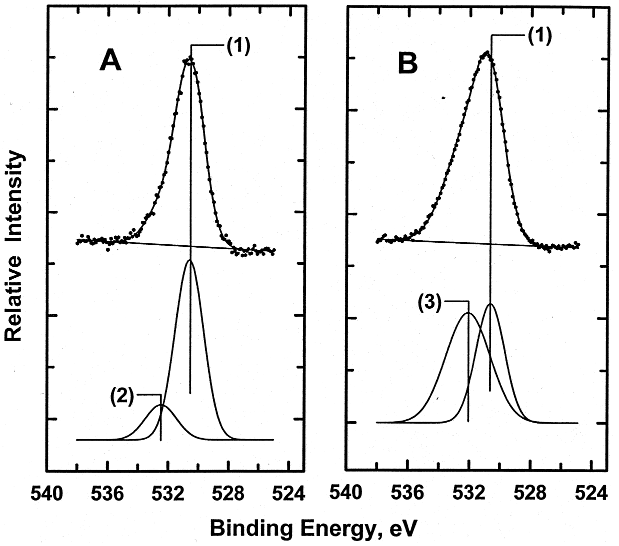


FIGURE 9 Deconvoluted O 1s XPS peaks of a TiO_2 substrate exposed to aqueous solution for (A) 10 min and (B) 12 hr. [Reproduced with permission from *Langmuir*, Vol. 15, No. 21, 1999, p. 7357; Copyright © 1999, American Chemical Society.]

10 min; spectrum "B" shows the same substrate after a 12-hr solution exposure. Deconvoluted peaks (1), (2), and (3) reveal different oxygen chemical environments arising from TiO_2, —OH, and H_2O, respectively. The contribution from —OH/H_2O increased with greater solution exposure time.

Chemical oxidation state identification from XAES features is accomplished using the concept of the *Auger parameter* (α), which is defined as the distance between the most prominent photoelectron emission and Auger features from the same element in an XP spectrum:

$$\alpha = E_k(\text{Auger}) - E_k(\text{PE}) \qquad (11)$$

where E_k(Auger) and E_k(PE) are kinetic energies of the Auger and photoelectrons, respectively. In order to avoid negative α values and since $E_k(\text{PE}) = h\nu - E_b(\text{PE})$, Eq. (11) can be rearranged to:

$$\alpha + h\nu = E_k(\text{Auger}) + E_b(\text{PE}) \qquad (12)$$

$$\alpha^* = E_k(\text{Auger}) + E_b(\text{PE}) \qquad (13)$$

Thus, the *modified* Auger parameter (α^*; AP) is independent of the photon energy and is always positive. Since the AP is determined from energy differences rather than absolute energy measurements, its value is not affected by surface charging. The AP is therefore often more sensitive to chemical state changes than the photoelectron E_b peak centers. The AP varies inversely with the photon energy, i.e., the higher the $h\nu$ value, the lower the AP and vice versa. These AP data are typically plotted in two dimensions (Fig. 10). Since the AP data utilizes

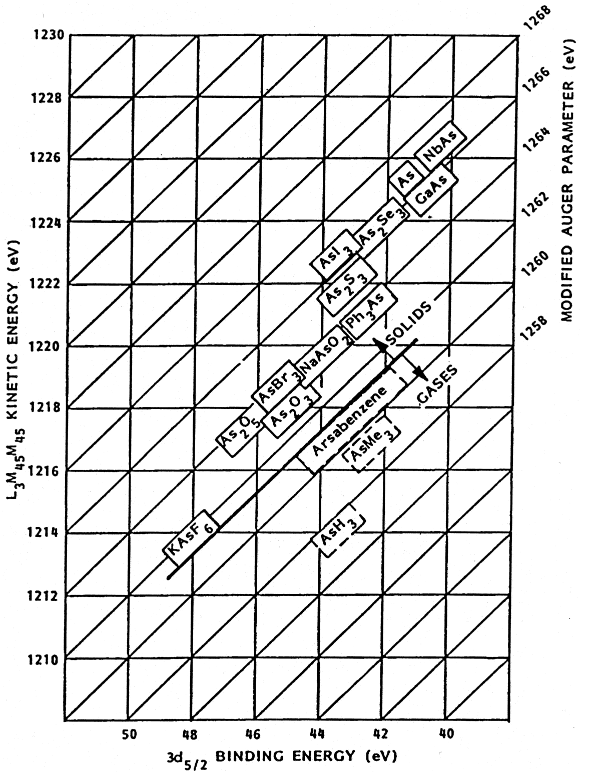


FIGURE 10 Modified Auger parameter plot. [Reprinted with permission from C. D. Wagner and A. Joshi, *Journal of Electron Spectroscopy and Related Phenomena*, Vol. 47, p. 301, 1988; Copyright © 1988, Elsevier Science.]

peak positions from both XPS and XAES, it is a particularly powerful means of differentiating various chemical oxidation states. The AP values are obtained via measured binding energy distances between the Auger transition (As $L_3M_{4,5}M_{4,5}$) and the core-level line (As $3d_{5/2}$). With this arrangement, the diagonal line represents equal sums of both Auger kinetic energy and XPS binding energy. The AP value can be obtained from the right-hand ordinate axis. It is noteworthy that while the original concepts of both α and α^* were empirical, they are directly related to extra-atomic relaxation energy (screening of outgoing Auger and photoelectron energy from neighboring atoms).

B. Quantitative Analysis

Assuming that the sample is homogeneous the number of photoelectrons per second (intensity) detected in an XP spectral peak is given by

$$I = n \cdot f \cdot \sigma \cdot \theta \cdot y \cdot \lambda \cdot A \cdot T \qquad (14)$$

where n = number of atoms per cm^3 of the analyte, f is the X-ray flux in photons/$cm^2 \cdot$ sec, σ = the photoelectric cross section for the atomic orbital of interest in cm^2,

θ = the angular efficiency factor for the instrument, y = the efficiency of the photoelectric process (i.e., the probability a photon causes the ejection of a photoelectron), λ is the IMFP, A = the area of the analyte sample in cm^2, and T = the detection frequency. Equation (14) can be rewritten as:

$$n = \frac{I}{f \cdot \sigma \cdot \theta \cdot y \cdot \lambda \cdot A \cdot T} \quad (15)$$

The denominator term is defined as the *atomic sensitivity factor*, S, which differs for various elemental photoelectron transitions. Equation (15) can then be further extended to a generalized expression for determining the mole fraction of any constituent element from the sum of the peak intensities from all elements in the sample being analyzed:

$$C_x = \frac{n_x}{\sum_i n_i} = \frac{\frac{I_x}{S_x}}{\sum_i \frac{I_i}{S_i}} \quad (16)$$

where C is the concentration of analyte, x, of interest being scanned. The element, x, is a subset of all elements, i. Thus, given the measured absolute intensities of different core levels along with their respective sensitivity factors, the atomic percent composition of elements on the surface can be computed. It should be noted that this formula assumes a homogeneous distribution of elements throughout the sample. The structural model of the analyte needs to be known to insure accurate determinations of concentration. For example, an oxide thin film covering the topmost layers of a refractory metal substrate will attenuate underlying photoelectron signal while accentuating the relative oxygen intensity. Thus, the atom percent determination of the oxygen in this system would be an overestimation if the oxide overlayer structure is not taken into account.

C. Initial and Final State Effects

Koopmans' theorem predicts that the XP spectra observed represents the electronic states of electrons in the analyte atom before the photoemission process, i.e., the *initial* state. *Spin orbit coupling* is an example of an *initial* state effect, in which spectral features arise from the inherent unpaired electron make-up of the atoms prior to the photoemission event. In this scenario, peak splittings due to energy differences between singlet and triplet states via interaction of the spin and orbital magnetic moments occur whenever there are unpaired electrons in the valence shells. If the electron ejected is parallel (triplet state) to that of the valence electrons, it can undergo exchange interaction and result in a lower kinetic energy (higher binding energy) than the case for an anti-parallel spin (singlet state). The result is a doublet in the XP spectrum. Splittings due to spin orbit coupling are generally not observed (resolved into 2 peaks) for low atomic number elements ($Z \leq 20$). A Ca 2p core level (Fig. 11), for instance, gives two photoemission peaks: $2p_{1/2}$ ($l = 1$ and $J = 1 - 1/2$) and $2p_{3/2}$ ($l = 1$ and $l = 1 + 1/2$). Similarly, d and f orbitals can split upon photoionization. The relative intensity ratio of the 2 peaks of a spin-orbit-coupled doublet is determined by the 2J + 1 multiplicity of the levels. For example, the relative intensity of J = 3/2 and J = 1/2 components of a 2p is 4:2 = 2:1; for that of 5/2 and 3/2 peaks of a 3d level, it is 6:4 and for the 7/2 and 5/2 peaks of the 4f level, it is 8:6.

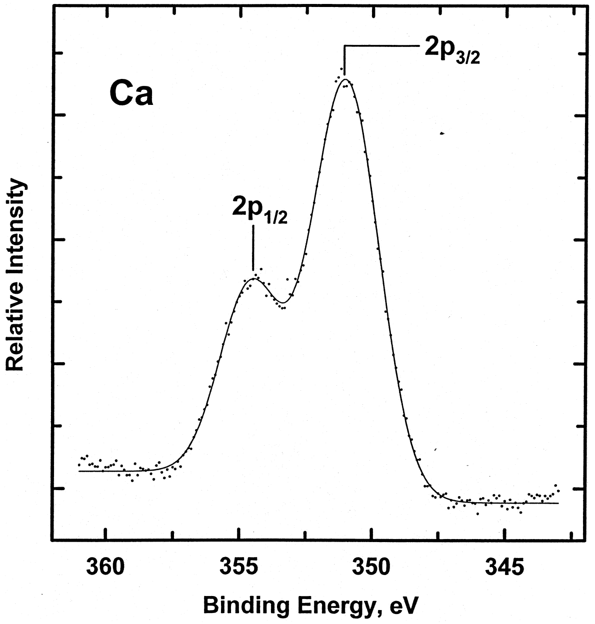


FIGURE 11 The XP spectrum of Ca 2p level showing spin-orbit-coupled $l \pm s = 1/2$ and 3/2 peaks.

Binding energy shifts from initial state effects can be interpreted using the charge potential model:

$$E_b^i = kq_i + \sum_j \frac{q_i}{r_{ij}} + E_b^{ref} \quad (17)$$

where E_b^i = the binding energy of an electron from an atom i, q_i is the charge on the atom, k = constant, q_j = the charge on a neighboring atom j, r_{ij} = the distance between atom i and atom j, and E_b^{ref} = the energy reference. The kq_i term indicates that binding energy increases with increasing positive charge on the atom from which the photoelectron emanates. The $\sum_j \frac{q_j}{r_{ij}}$ term, known as the *Madelung* sum, in ionic solids negates the contribution from the kq_i term, since the neighboring atom has an opposite charge.

During the photoionization process, changes in the electronic environment due to the creation of the core-level vacancy leading to *final* state effects play a large role in influencing the binding energy. In order to properly

interpret the XPS binding energy value, perturbation of the electronic environment during the photoemission must be accounted for. For instance, let $N =$ the number of total electrons in the atom before photoionization. The atom in its initial state conditions E_N^i absorbs a monochromatic photon of energy, $h\nu$, causing the ejection of a photoelectron with kinetic energy, E_k. The adsorption process takes place in approximately 10^{-17} sec. Approximately 10^{-14} sec later, the atom itself has one less electron and a core-level vacancy. The energy balance between initial and final states of the atom before and after photoionization can be expressed as:

$$E_N^i + h\nu = E_{N-1,l}^f + E_k \tag{18}$$

where $E_N^i =$ total energy of the atom with N electrons in the initial state (i.e., before photoionization), $E_{N-1,l}^f =$ the total energy of the atom with N-1 electrons and a hole in the core level, l, in the final state. It should be emphasized that the $N - 1$ remaining electrons in the final state atom and electrons in neighboring atoms are influenced by the presence of the core shell vacancy, relaxing to lower the total energy of the of the atom by ΔE_{relax}. The relaxation energy should be accounted for in the determination of the kinetic energy of the photoelectron. The binding energy is not simply the energy of the orbital from which the photoelectron is emitted (i.e., initial state effect), but rather the difference in energy resulting from the perturbation of the remaining electrons upon removal of a core level electron. These photoemission processes occur at time scales sufficiently slow to influence exiting electrons via attraction of the core-ionized atom, known as the *adiabatic limit*. In the other extreme, photoelectrons can be emitted before the core-ionized atom relaxes. Photoemission during these "fast" processes (known as the *sudden limit*) often result in extra peaks in the XP spectrum. *Shake-up* and *shake-off* losses are final state effects which appear in the XP spectrum resulting from a photoelectron imparting energy to another electron within the atom. These features arise from the perturbation process (final state effects) caused by photoemission. The energy associated with relaxation may be sufficient to excite a valence level electron to higher energy. The electron receiving energy either ends up in a higher unoccupied state having discrete energy (shake-up) or an unbounded state (shake-off). Since photoemission and relaxation occur simultaneously, the outgoing photoelectron loses kinetic energy. These shake peaks, due to kinetic energy losses, appear at higher binding energy relative to the main core-level peak. Discrete shake-up losses are pronounced for metal oxides. Pronounced intensities are typically found for compounds having unpaired 3d or 4f electrons. Shake-up features that show up in the XP spectra of the Cu 2p core levels in Cu oxide stand out as a notable example (Fig. 12), providing noteworthy diagnostic tool for detecting the Cu^{2+}. The CuO spectrum shows pronounced shake-up satellite features (top) due to the fact that Cu^{2+} ([Ar]$3d^9$) has an open shell configuration. In contrast, Cu_2O (bottom) lacks these features since Cu^+ ([Ar]$3d^{10}$) has a closed shell arrangement.

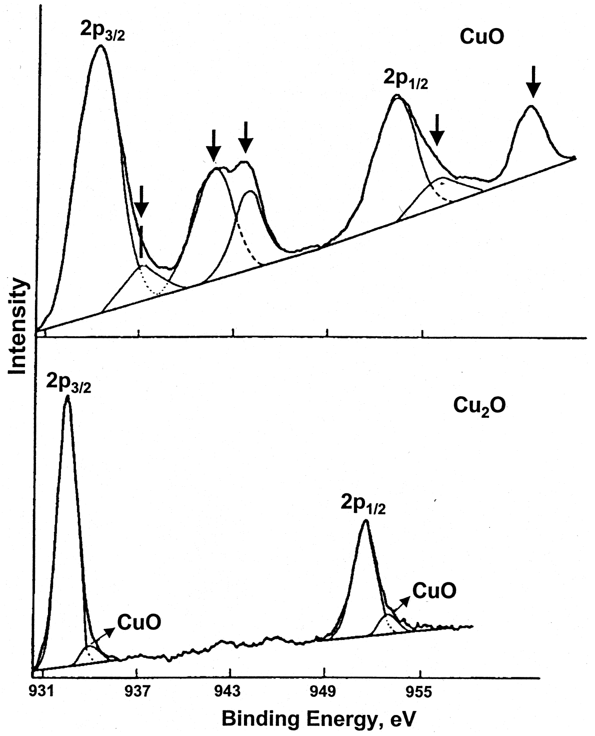


FIGURE 12 The XP scans of Cu 2p core levels of CuO (top) with pronounced shake-up satellites. [Reprinted from *Chemical Physics Letters*, Vol. 63, M. Scrocco, p. 53, Copyright © 1979, with permission from Elsevier Science.]

IV. EXAMPLES

The number of examples showing the utility of XPS, widely cited under many research journal titles, are too numerous to be included in a single encyclopedia article. Thus, only two illustrations are presented in this section with specific emphasis on applications for surface science studies to stimulate the reader's interest. Extensive reviews on the latest developments in the technique have been regularly published bi-annually on even-numbered years in *Analytical Chemistry*, a journal publication of the American Chemical Society. The *Journal of Electron Spectroscopy and Related Phenomena*, published by Elsevier, Amsterdam, is specifically devoted to the field of XPS as well as other electron spectroscopies. The first example addresses the use of XPS for characterizing the electronic

structure of the metal—metal bond in bimetallic systems; the second illustration is on the use of XPS to indirectly measure nanoscale CuO particle sizes.

A. Electron Donor-Electron Acceptor Interactions in Bimetallic Systems

In recent years, the electronic, chemical, and catalytic properties of bimetallic surfaces generated by vapor-depositing one metal onto a crystal face of a second metal has been the focus of considerable attention, motivated by the roles that bimetallic surfaces play in catalysis, electrochemistry, and microelectronics. Core-level binding energy shifts in bimetallic (transition metal) systems show that in the formation of the surface metal—metal bond, a gain in electrons by the element initially having a larger fraction of empty states in the valence band occurs. The trend observed, however, contradicts predictions based on the relative electronegativities of the bulk alloys. XPS thus shows that adsorbed surface metal atoms, called *adatoms*, have a different intrinsic electronic nature compared to bulk metal.

In the experiment, 1.0 monolayer (ML) coverages of Pd and Cu metal were deposited onto various single crystal transition metals via "hot" filament evaporation, followed by high-resolution XP scans. Figure 13 (top) shows the Pd $3d_{5/2}$ binding energy differences between the surface adatoms and that of bulk Pd for various supports. In all cases, the direction of the electronic perturbation for the Pd adatoms can be characterized with a model describing charge transfer from the Pd overlayer into the substrate. The magnitude of the electronic perturbation increases as the transition-metal substrate "moves" from right to left in the periodic table. For pseudomorphic Pd adatoms, the surface atomic density follows the order Ta(110), 1.30×10^{15} atoms/cm^2 < W(110), 1.42×10^{15} atoms/cm^2 ≈ Mo(110), 1.43×10^{15} atoms/cm^2 < Re(0001), 1.54×10^{15} atoms/cm^2 < Ru(0001), 1.57×10^{15} atoms/cm^2. Bimetallic systems with larger Pd—Pd bond distances (i.e., weaker Pd—Pd interactions) have stronger Pd-substrate interactions, thus resulting in larger electronic perturbations. In the case of Cu adatoms (Fig. 13, bottom), the binding energies are shifted either positively or negatively with respect to the bulk Cu value depending on the support. The bulk value for Cu was obtained from scans of the top layer of a Cu(100) crystal. The XPS data are consistent with a model in which Cu atoms supported on metals on the right side of the periodic table have electron densities larger than that of Cu(100) surface atoms. Similarly, a reverse phenomenon is observed for Cu atoms supported on metals on the left side of the periodic table. For pseudomorphic Cu monolayers, the surface atomic density

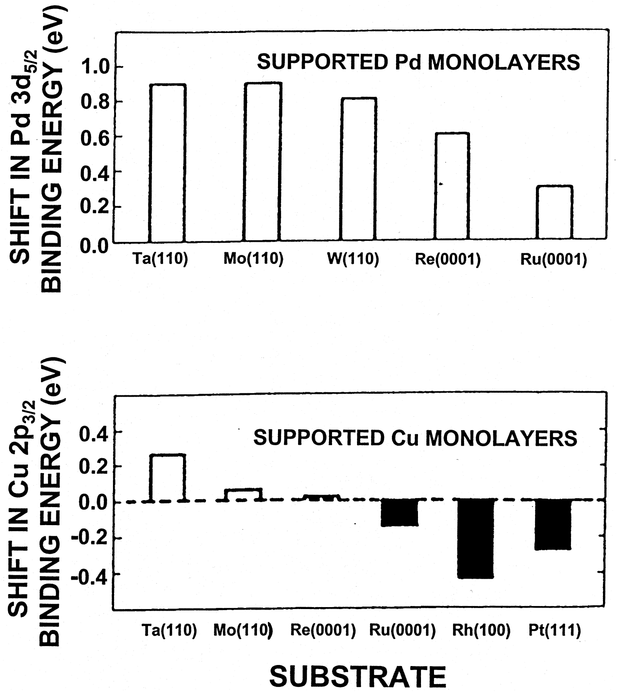


FIGURE 13 Top: Difference in Pd $3d_{5/2}$ XPS binding energy between 1.0 ML Pd and surface atoms of Pd(100) as a function of metal substrate. Bottom: Difference in Cu $2p_{3/2}$ XPS binding energy between 1.0 ML Cu and surface atoms of Cu(100) as a function of metal substrate. [Reproduced with permission from *The Journal of Physical Chemistry*, Vol. 95, No. 15, 1991; Copyright © 1991, American Chemical Society.]

follows the sequence: Ta(110) < Rh(100), 1.38×10^{15} atoms/cm^2 < Mo(110) < Pt(111), 1.51×10^{15} atoms/cm^2 ≈ Re(0001) < Ru(0001) (Fig. 13, bottom). Thus, similar to Pd, bimetallic systems with the larger Cu—Cu adatom bond distances have the larger overlayer-substrate interactions.

Trends seen for both Pd and Cu adlayers can be explained by a simple model that takes into account the type of metals present. The strongest electronic perturbations are found for systems that involve a combination of a metal with an almost fully occupied valence band and a metal with a valence band more than half empty. The formation of a surface metal—metal bond generally leads to a gain in electron density by the element initially having the larger fraction of empty states in its valence band. The direction of electron transfer can be easily understood in terms of orbital mixing: hybridization of the occupied states of an electron-rich metal A with the unoccupied levels of an electron-deficient metal B. This mixing leads to a loss of A character in the occupied states and hence a reduction in the electron density on metal A. In the case of Pd (an electron-rich admetal), the relative core-level shift of the 1.0 ML of the Pd $3d_{5/2}$ level (with respect to the

bulk value) supported on various transition metals similarly follows the trend: Ta > W > Re > Ru. The core-level shift decreases in the left-to-right direction of the periodic table. The magnitude of the perturbations induced by the loss of electron density increases as the fraction of empty levels in the valence band of the metal substrate increases: Ta > W > Re > Ru (Fig. 14a). Cu, on the other hand, has a 4s valence band that is half empty and can act as either an electron donor or acceptor, depending upon the fraction of empty states in the valence band of the metal substrate.

Interestingly, there is an excellent correlation between the changes in the admetal binding energies and the relative abilities of these transition-metal substrates to adsorb Pd as measured by temperature programmed desorption (TPD), which can also be explained using the model described above. [The TPD technique involves adsorbing the analyte metal onto a substrate followed by linearly heating the sample to desorb it and subsequent detection via mass spectrometer. A plot of the ion current (mass spectral intensity) is obtained as a function of temperature. The peak temperature maximum of the resulting spectrum is indicative of the admetal binding strength to the surface.] Desorption temperatures of Pd monolayers from several surfaces are presented in Fig. 14b. These data show a general trend of stronger Pd-substrate bond strengths for elements with the least occupied valence band. The Pd/Ta(110) system has a desorption temperature of 1540 K, the highest observed, and the Pd/Ru(0001) has the lowest desorption temperature (1440 K). The larger the electronic perturbations on the Pd atoms, the stronger the bimetallic bond. An interrelationship between the magnitude of the Pd $3d_{5/2}$ core level shift and the Pd-substrate binding strength is thus readily observed. Similarly, a strong correlation has been found between changes in the XPS core-level binding energies and variations in the ability for the supported films to adsorb CO. Temperature programmed desorption data indicate CO desorption temperatures from $Pd_{1.0\,ML}$/Ta(110) ($\sim$250 K), $Pd_{1.0\,ML}$/W(110) ($\sim$300 K), $Pd_{1.0\,ML}$/Ru(0001) ($\sim$330 K) that are much lower than that from Pd(100) ($\sim$480 K). According to the above-described model, the electron density of supported monolayers of Pd is smaller than that of the surface atoms of Pd(100). A partial positive charge on the Pd adatoms is consistent with a reduction in their ability to coordinate CO via π–back–bonding, producing a weaker Pd—CO bond on the supported monolayers compared to Pd(100). The model described above predicts that the Pd surface atom electron density and ability to π–back–donate will follow the sequence $Pd_{1.0\,ML}$/Ta(110) < $Pd_{1.0\,ML}$/W(110) < $Pd_{1.0\,ML}$/Ru(0001) < Pd(100). This exact trend is exhibited in the TPD data. For monolayers of Cu (having a relatively electron-deficient density) deposited on electron-rich metals (Ru, Rh, and Pt), the model predicts an enhancement in electron density and π–back–bonding capacity of the Cu adatoms with respect to the surface atoms of Cu(100). Adsorption of CO on Cu films induces a large increase in the Cu $2p_{3/2}$ XPS binding energy, in part, due to π–back–donation. Figure 15 illustrates how the CO-induced shift in the Cu $2p_{3/2}$ peak position (measured after saturating the Cu surfaces with CO at 100 K) changes with the metal substrate. The CO—induced shift increases as the metal substrate "moves" from left to right in the periodic table. This trend is attributed to an increase in the ability of the Cu adatoms to π–back–donate.

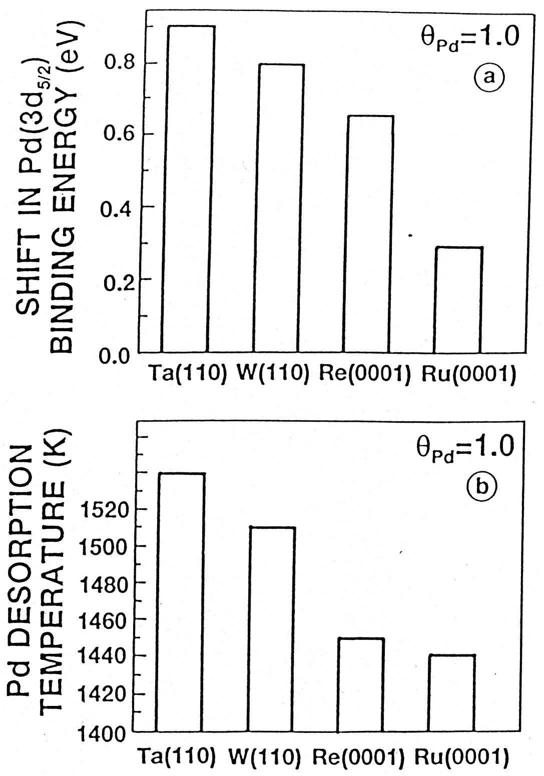


FIGURE 14 (a) Binding energy shift in Pd $3d_{5/2}$ level for 1.0 ML Pd on several single crystal substrates relative to the surface atoms of Pd(100). (b) The TPD desorption peak temperature maxima of 1.0 ML Pd deposited on several substrates. [Reprinted with permission from *Physical Review B*, Vol. 46, No. 1, p. 7082, 1992; Copyright © 1992, American Physical Society.]

B. Finite Size Effects on Core-Level-to-Shake-up Satellite Intensities

Surface characterization of CuO particles supported on SiO_2 is a catalyst system of importance for industrial applications, such as the synthesis of methanol. X-ray photoelectron spectroscopy has been shown to be an

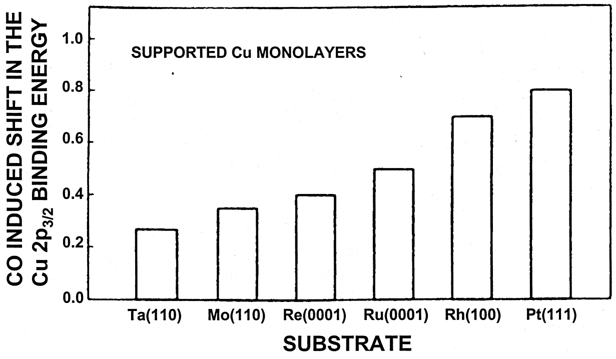


FIGURE 15 CO-induced XPS binding energy shifts of the Cu $2p_{3/2}$ level measured after saturation the 1.0 ML Cu/metal substrate surfaces with CO at 100 K. [Reproduced with permission from *The Journal of Physical Chemistry*, Vol. 95, No. 15, 1991; Copyright © 1991, American Chemical Society.]

effective tool for indirectly measuring CuO particle size in the Cu/SiO_2 system.

In the experiment, the CuO/SiO_2 model catalyst was prepared by depositing $Cu(CH_3CO_2)_2{\bullet}H_2O$ [$Cu(ac)_2$] dissolved in butanol onto an SiO_2 support via spin-coating. The SiO_2 substrate with this adlayer was then calcinated to 450°C to remove unwanted organic substituents and hence form spherically shaped CuO particles. The particle sizes were varied via changing the $Cu(ac)_2$ solution concentrations. Four size distributions were produced with mean particle heights (corresponding to mean diameter) of 3.7, 4.1, 4.4, and 6.3 nm, formed from concentrations of 0.0040 M, 0.0070 M, 0.0085 M, and 0.010 M $Cu(ac)_2$, respectively. These dimensions were verified using atomic force microscopy (AFM).

Figure 16A shows a stackplot of XP spectra of the Cu 2p core regions acquired from these surfaces after AFM imaging. Shake-up features at ~945 and ~965 eV for the Cu $2p_{3/2}$ and $2p_{1/2}$ core levels are evident and are diagnostic of an open $3d^9$ shell of Cu(+2). The peak positions and relative intensities of the satellites from these levels are indicative of the presence of CuO at the surface. The relative intensities of the shake-up lines to the main core level of both the Cu $2p_{3/2}$ and $2p_{1/2}$ levels varied as a function of $Cu(ac)_2$ solution concentration. The shake-up intensities denoting CuO on the surface were relatively more intense at higher $Cu(ac)_2$ concentration.

The peakfit of the Cu $2p_{3/2}$ core level revealed two binding energy states (with fwhm in parentheses) at 932.8 (1.91) and 933.8 (3.12) eV, which we assign to a Cu(0/+1) state and CuO, respectively. The binding energy region scanned to obtain these Cu 2p peaks (925–975 eV) took approximately 40 min to acquire. It was during this acquisition time that X-ray irradiation from XPS caused reduction of the CuO particles. For the smaller CuO particles, the ratio of exposed surface area to bulk is greater, which

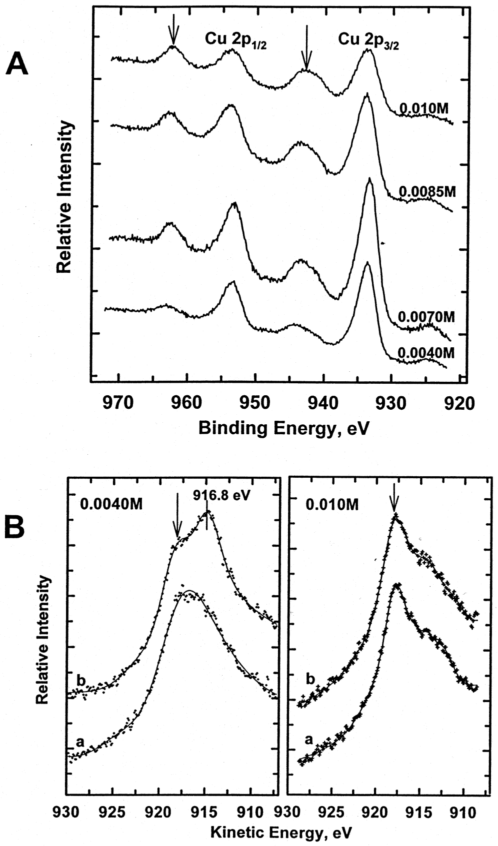


FIGURE 16 (A) The XPS of Cu $2p_{1/2}$ and $2p_{3/2}$ peaks of varying particle sizes of CuO on SiO_2 prepared from various $Cu(ac)_2$ concentrations; (B) The XAES of Cu LMM spectra upon (a) immediate XP scans and (b) after a 50 min X-ray exposure. [Reproduced with permission from *Langmuir*, Vol. 15, No. 8, 1999, p. 2807; Copyright © 1999, American Chemical Society.]

results in an overall increased dosage of X-ray irradiation and hence greater susceptibility to reduction. In addition, the presence of adventitious carbon obtained from treating these substrates in air likely enhanced reduction. From our XPS measurements of the C 1s level intensities and taking into account atomic sensitivity factors for all of the orbitals scanned (including O 1s, Cu $2p_{3/2}$ and Si 2p), there was 12–25 atom percent carbon on these surfaces. Further, XPS induced reduction of Cu(+2) to Cu(+1) has been reported to increase in the presence of carbonaceous overlayers.

To test this hypothesis of particle size dependent reduction rates, the Cu $L_3M_{4,5}M_{4,5}$ XAES regions were examined before and after exposure of the substrates to the soft X-rays. Figure 16B shows Auger spectra of the Cu $L_3M_{4,5}M_{4,5}$ transition of 0.010 M and 0.0040 M $Cu(ac)_2$ prepared surfaces corresponding to the largest and smallest CuO particle sizes, respectively. Spectra "a" of both samples were taken immediately after initial X-ray irradiation and "b" after 50 additional min of exposure to

the X-ray gun. Instrumental parameters and X-ray exposure times were identical for both particle sizes scanned. The arrows at 917.8 eV denote the XAES peak position of CuO, which is consistent with the literature. No difference in the Auger line shape or position was seen in the 0.010 *M* $Cu(ac)_2$ prepared sample; however, a marked difference was seen for the 0.0040 *M* substrate before and after the 50-min X-ray exposure. Even upon initial exposure to the X-ray source, a difference in the Auger line shapes can be seen between the two particle sizes in spectra "a". This difference in the Auger line shape between spectrum "a" of the 0.0040 *M* $Cu(ac)_2$ and that of 0.010 *M* $Cu(ac)_2$ indicates that some reduction has already taken place during this initial scan. After 50 min, a feature at 916.8 eV in spectrum "b" of the 0.0040 *M* $Cu(ac)_2$ substrate, which is assigned to a Cu(+1) state (Cu_2O), dominates. Some intensity within this spectral region can be seen in "a" but is relatively weak compared to the 917.8 eV position. No intensity was observed within the 918.8 eV Cu $L_3M_{4,5}M_{4,5}$ region corresponding to that reported for metallic Cu(0). It can thus be concluded that Cu(0) is not present on the surface and that the CuO particles are likely reduced to Cu_2O. Clearly, there is greater reduction for the smaller particles.

Figure 17A shows a representative peakfit of the Cu $2p_{3/2}$ core level and its corresponding shake-up satellites. This particular fit is for the 0.0040 *M* $Cu(ac)_2$ treated sample that exhibited the largest core-level intensity due to Cu(+1). Gaussian line shapes with a linear background were used to fit peak 1, which is assigned to the Cu(+1) state; peaks 2, 3, and 4 are assigned to the Cu(+2) state. The ratio of the sum of the areas of peaks 2, 3, and 4 to that of peak 1 was used as a measure of the relative amount of Cu(+2) to Cu(+1) on the surface. As the Cu oxide particle size increases, the relative amount of XPS-induced reduction decreases. Intensity from Cu(+2) dominated in XP scans of larger particles (6.3 nm) while intensity from Cu(+1) dominated in XP scans of smaller ones (3.7 nm); in addition, there was a decrease in the Cu 2p shake-up intensity (Fig. 16A). Figure 17B shows a plot of $[Cu^{+2}]/[Cu^{+}]$ calculated from the above-mentioned peak areas as a function of the AFM-measured cluster heights of the CuO particles; $[Cu^{2+}]/[Cu^{+}]$ varies linearly and increases with particle size.

These examples represent only a glimpse into the numerous applications of XPS to obtain information (both qualitative and quantitative) regarding the chemical oxidation state, atomic composition, and electronic structure of surfaces. When used in combination with complementary surface analytical probes (such as TPD and AFM as illustrated in the preceding examples), XPS can be an especially powerful technique for obtaining a detailed picture of the solid-solid/solid-gas interface. For further examples and discussion of the broad scope of applications offered by XPS, the reader is referred to the supporting literature cited in the captions of the figures and bibliography section.

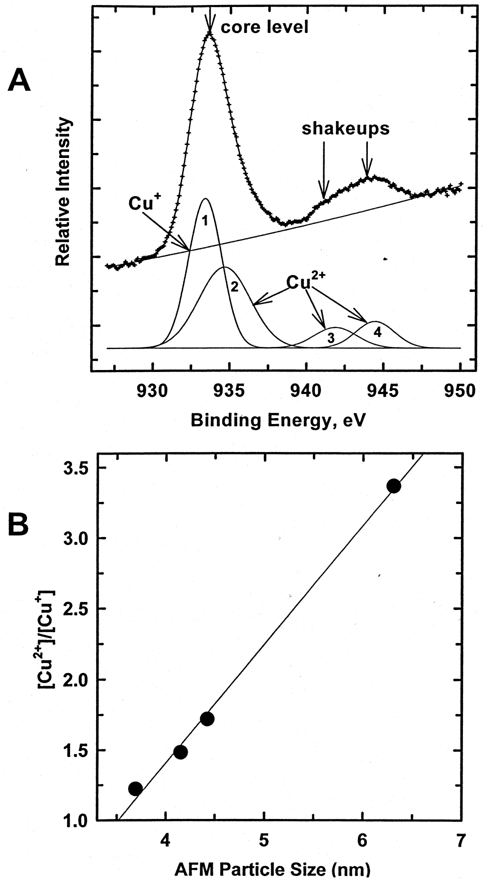


FIGURE 17 (A) Curvefit of XPS Cu $2p_{3/2}$ core level. Peak 1 denotes the binding energy state for Cu(+1). Peaks 2–4 denote the Cu(+2) state; (B) Plot of $[Cu^{2+}]/[Cu^{+}]$ from XPS Cu $2p_{3/2}$ peak areas as a function of particle size using AFM measured cluster height. [Reproduced with permission from *Langmuir*, Vol. 15, No. 8, 1999, p. 2808; Copyright © 1999, American Chemical Society.]

SEE ALSO THE FOLLOWING ARTICLES

Auger Electron Spectroscopy • Molecular Electronics • Photoelectron Spectroscopy • Surface Chemistry • Vacuum Technology • X-Ray Analysis • X-Ray, Synchrotron Radiation and Neutron Diffraction

BIBLIOGRAPHY

Barr, T. L., and Seal, S. (1995). Nature and use of adventitious carbon as a binding energy standard. *J. Vacuum Sci. Technol. A* **13,** 1239–1246.

Briggs, D. (1998). "Surface Analysis of Polymers by XPS and Static SIMS," Cambridge Univ. Press, Cambridge, U.K.

Henrich, V. E., and Cox, P. A. (1996). "The Surface Science of Metal Oxides," Cambridge Univ. Press, Cambridge, U.K.

Hüfner, S. (1996). "Photoelectron Spectroscopy: Principles and Applications, 2nd edition," Springer-Verlag, Berlin.

Oleford, I. (1997). X-ray photoelectron spectroscopy. *In* "Surface Characterization: A User's Sourcebook" (D. Brune, R. Hellborg, H. J. Whitlow, and O. Hunderi, eds.), pp. 291–319. Wiley-VCH Verlag, Weinheim, Germany.

Powell, C. J., Jablonski, A., Tilinin, I. S., Tanuma, S., and Penn, D. R. (1999). Surface sensitivity of Auger-electron spectroscopy and X-ray photoelectron spectroscopy. *J. Electron Spectrosc. Related Phenomena* **98–99,** 1–15.

Somorjai, G. A. (1994). "Introduction to Surface Chemistry and Catalysis," Wiley, New York.

Turner, N. H., and Schreifels, J. A. (2000). Surface analysis: X-ray photoelectron spectroscopy and Auger electron spectroscopy. *Anal. Chem.* **72,** 99R–110R.

Woodruff, D. P., and Delchar, T. A. (1994). "Modern Techniques of Surface Science, 2nd edition," Cambridge Univ. Press, Cambridge, U.K.